Behavioral and Brain Sciences (forthcoming)

This Target Article has been accepted for publication and has not yet been copyedited and proofread. The article may be cited using its doi (About doi), but it must be made clear that it is not the final version.

Deep Problems with Neural Network Models of Human Vision

Jeffrey S. Bowers[1], j.bowers@bristol.ac.uk; https://jeffbowers.blogs.bristol.ac.uk/

Gaurav Malhotra[1], gaurav.malhotra@bristol.ac.uk

Marin Dujmović[1], marin.dujmovic@bristol.ac.uk

Milton Llera Montero[1], m.lleramontero@bristol.ac.uk

Christian Tsvetkov[1], christian.tsvetkov@bristol.ac.uk

Valerio Biscione[1], valerio.biscione@gmail.com

Guillermo Puebla[1], guillermo.puebla@bristol.ac.uk

Federico Adolfi[1,2], fedeadolfi@gmail.com

John E. Hummel[3], jehummel@illinois.edu

Rachel F. Heaton[3], rmflood2@illinois.edu

Benjamin D. Evans[4], b.d.evans@sussex.ac.uk

Jeffrey Mitchell[4], j.mitchell@napier.ac.uk

Ryan Blything[5], r.blything@aston.ac.uk

[1] School of Psychological Science, University of Bristol, UK; [2] Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Germany; [3] Department of Psychology, University of Illinois Urbana-Champaign, USA; [4] Department of Informatics, School of Engineering and Informatics, University of Sussex, UK.[5]

Short Abstract: Deep neural networks (DNNs) are often described as the best models of biological vision based on their successes in predicting behavioral and brain responses to images of objects. We show that these good predictions may be mediated by DNNs that share little overlap with biological vision and that these same DNNs account for almost no results from psychological research. We argue that theorists interested in developing biologically plausible models of human vision need to direct their attention to explaining psychological findings, and more generally, build models that explain the results of experiments that manipulate independent variables designed to test hypotheses.

Long Abstract: Deep neural networks (DNNs) have had extraordinary successes in classifying photographic images of objects and are often described as the best models of biological vision. This conclusion is largely based on three sets of findings: (1) DNNs are more accurate than any other model in classifying images taken from various datasets, (2) DNNs do the best job in predicting the pattern of human errors in classifying objects taken from various behavioral datasets, and (3) DNNs do the best job in predicting brain signals in response to images taken from various brain datasets (e.g., single cell responses or fMRI data). However, these behavioral and brain datasets do not test hypotheses regarding what features are contributing to good predictions and we show that the predictions may be mediated by DNNs that share little overlap with biological vision. More problematically, we show that DNNs account for almost no results from psychological research. This contradicts the common claim that DNNs are good, let alone the best, models of human object recognition. We argue that theorists interested in developing biologically plausible models of human vision need to direct their attention to explaining psychological findings. More generally, theorists need to build models that explain the results of experiments that manipulate independent variables designed to test hypotheses rather than compete on making the best predictions. We conclude by briefly summarizing various promising modelling approaches that focus on psychological data.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

Key Words: Brain-Score; Computational Neuroscience; Deep Neural Networks; Human Vision; Object recognition.

Word counts: Abstract: 232; Main Text:14,990; References: 4945; Entire Text: 20,478

1 Introduction

The psychology of human vision has a long research history. Classic studies in color perception (Young, 1802), object recognition (Lissauer, 1890), and perceptual organization (Wertheimer, 1912) date back well over 100 years, and there are now large and rich literatures in cognitive psychology, neuropsychology, and psychophysics exploring a wide range of high- and low-level visual capacities, from visual reasoning on the one hand to subtle perceptual discriminations on the other. Along with rich datasets there are theories and computational models of various aspects of vision, including object recognition (e.g., Biederman, 1987; Cao, Grossberg, & Markowitz, 2011; Marr, 1982; Erdogan & Jacobs, 2017; Hummel & Biederman, 1992; Ullman & Basri, 1991; for reviews see Gauthier & Tarr, 2016; Hummel, 2013). However, one notable feature of psychological models of vision is that they typically do not solve many engineering challenges. For example, the models developed in psychology cannot identify naturalistic images of objects.

By contrast, deep neural networks (DNNs) first developed in computer science have had extraordinary success in classifying naturalistic images and now exceed human performance on some object recognition benchmarks. For example, the *ImageNet Large Scale Visual Recognition Challenge* was an annual competition that assessed how well models could classify images into one of a thousand categories of objects taken from a dataset of over 1 million photographs. The competition ended in 2017 when 29 of 38 competing teams had greater than 95 percent accuracy, matching or surpassing human performance on the same dataset. These successes have raised questions as to whether the models work like human vision, with many researchers highlighting the similarity between the two systems, and some claiming that DNNs are currently the best models of human visual object processing (e.g., Kubilius et al., 2019; Mehrer, et al, 2021; Zhuanga et al., 2021).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

Strikingly, however, claims regarding the similarity of DNNs to human vision are made with little or no reference to the rich body of empirical data on human visual perception. Indeed, researchers in psychology and computer science often adopt very different criteria for assessing models of human vision. Here we highlight how the common failure to consider the vast set of findings and methods from psychology has impeded progress in developing human-like models of vision. It has also led to researchers making far too strong claims regarding the successes of DNNs in modelling human object recognition. In fact, current deep network models account for almost no findings reported in psychology. In our view, a plausible model of human object recognition must minimally account for the core properties of human vision.

The article is organized as follows. First, we review and criticize the main sources of evidence that have been used to support the claim that DNNs are the best models of human object recognition, namely, their success in predicting the data from a set of behavioral and brain studies. We show that good performance on these datasets is obtained by models that bear little relation to human vision. Second, we question a core theoretical assumption that motivates much of this research program, namely, the hypothesis that the human visual system has been optimized to classify objects. Third, we assess how well DNNs account for a wide range of psychological findings in vision. In almost all cases these studies highlight profound discrepancies between DNNs and humans. Fourth we briefly note how similar issues apply to other domains in which DNNs are compared to humans. Fifth we briefly outline more promising modeling agendas before concluding.

We draw two general conclusions. First, current DNNs are not good (let alone the best) models of human object recognition. Apart from the fact that DNNs account for almost no findings from psychology, researchers rarely consider alternative theories and models that do account for many key experimental results. Second, we argue that researchers interested in developing human-like DNN models of object

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

recognition should focus on accounting for key experimental results reported in psychology rather than the current focus on predictions that drive so much current research.

2 The problem with benchmarks

It is frequently claimed that DNNs are the best models of the human visual system, with quotes like:

> "Deep convolutional artificial neural networks (ANNs) are the leading class of candidate models of the mechanisms of visual processing in the primate ventral stream".
> Kubilius et al. (2019).

> "Deep neural networks provide the current best models of visual information processing in the primate brain" (Mehrer, et al, 2021).

> "Primates show remarkable ability to recognize objects. This ability is achieved by their ventral visual stream, multiple hierarchically interconnected brain areas. The best quantitative models of these areas are deep neural networks..." (Zhuanga et al., 2021).

> "Deep neural networks (DNNs) trained on object recognition provide the best current models of high-level visual areas in the brain..." (Storrs et al., 2021)

Relatedly, DNNs are claimed to provide important insights into how humans identify objects:

> "Recently, specific feed-forward deep convolutional artificial neural networks (ANNs) models have dramatically advanced our quantitative understanding of the neural mechanisms underlying primate core object recognition" (Rajalingham et al., 2018).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

And more generally:

> "Many recent findings suggest that deep learning can inform our theories of the brain…many well-known behavioral and neurophysiological phenomena, including… visual illusions and apparent model-based reasoning, have been shown to emerge in deep ANNs trained on tasks similar to those solved by animals". (Richards et al., 2020)

> "AI is now increasingly being employed as a tool for neuroscience research and is transforming our understanding of brain functions. In particular, deep learning has been used to model how convolutional layers and recurrent connections in the brain's cerebral cortex control important functions, including visual processing, memory, and motor control" (Macpherson et al., 2021).

Of course, these same authors also note that DNNs are still far from perfect models of human vision and object recognition, but it is the correspondences that are emphasized and that receive all the attention.

The claim that DNNs are the best models of human object recognition is largely justified based on three sets of findings, namely, (1) DNNs are more accurate than any other model in classifying images taken from various datasets, (2) DNNs do the best job in predicting the pattern of human errors in classifying objects taken from various behavioral studies, and (3) DNNs do the best job in predicting brain recordings (e.g., single-cell responses or fMRI bold signals) in response to images taken from various studies. According to this research program, all else being equal, the better the models perform on the behavioral and brain datasets the closer their correspondence with human vision. This is nicely summarized by Schrimpf et al. (2020a) when describing their benchmark dataset: "Brain-Score - a composite of multiple

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

neural and behavioral benchmarks that score any [artificial neural network] on how similar it is to the brain's mechanisms for core object recognition" (p 1).

A key feature of these behavioral and brain studies is that they assess how well DNNs predict behavioral and brain responses to stimuli that vary along multiple dimensions (e.g., image category, size, color, texture, orientation, etc.) and there is no attempt to test specific hypotheses regarding what features are contributing to good predictions. Rather, models are assessed and compared in terms of their predictions on these datasets after averaging over all forms of stimulus variation. For lack of a better name, we will use the term *prediction-based experiments* to describe this method. This contrasts with *controlled experiments* in which the researcher tests hypotheses about the natural world by selectively manipulating independent variables and comparing the results across conditions to draw conclusions. In the case of studying human vision, this will often take the form of manipulating the images to test a specific hypothesis. For instance, a researcher might compare how well participants identify photographs vs. line drawings of the same objects under the same viewing conditions to assess the role of shape vs. texture/color in object identification (see Section 4.2.3).

To illustrate the prediction-based nature of these studies consider the image dataset from Kiani et al. (2007) used by Khaligh-Razavi and Kriegeskorte (2014) to assess how well DNNs can predict single-cell responses in macaques and fRMI bold signals in humans using Representational Similarity Analysis (RSA). This dataset includes objects from six categories (See Figure 1) that vary in multiple ways from one another (both within and between categories) and the objects often contain multiple different visual features diagnostic of their category (e.g., faces not only share shape they tend to share color and texture). Critical for present purposes, there is no manipulation of the images to test which visual features are used for object recognition in DNNs, humans, or macaques, and what visual features DNNs use to support good predictions on the behavioral and brain datasets. Instead, models receive an overall RSA score that is used

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

to make inferences regarding the similarity of DNNs to the human (or macaque) visual object recognition system.

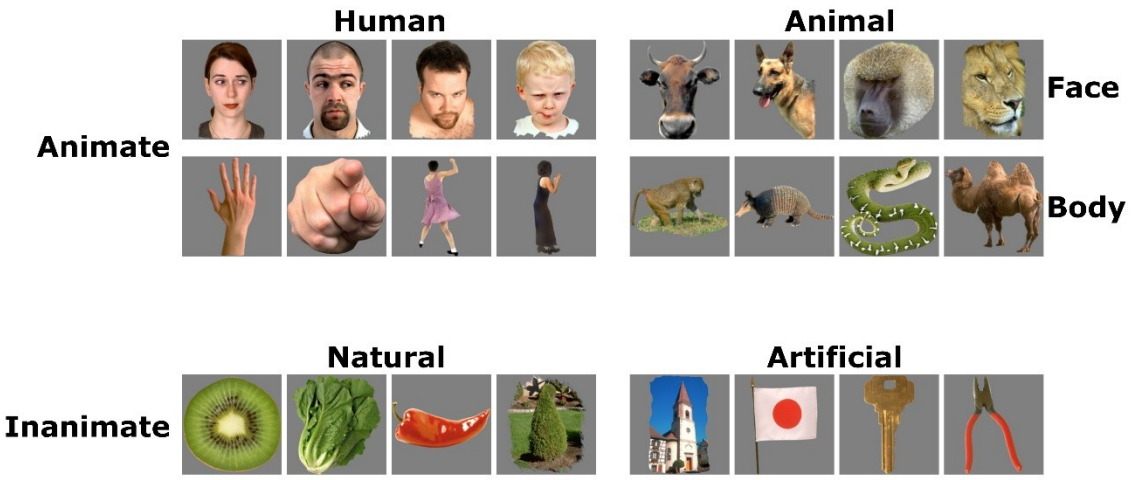


Figure 1. Example images from Kiani et al. (2007) that include images from six categories.

Or consider the Brain-Score benchmark that includes a range of behavioral and brain studies that together are used to rate a model's similarity to human object recognition (Schrimpf et al., 2020ab). Currently five studies are used to assess how well DNNs predict brain activation in inferotemporal (IT) cortex. The first of these (Majaj et al., 2015) recorded from neurons from two awake behaving rhesus macaques who viewed thousands of images when objects were placed on unrelated backgrounds with the size, position, and orientation of the objects systematically varied to generate a large dataset of images. See Figure 2 for some example images. Despite the manipulation of size, position, and orientation of the images, Brain-Score collapses over these factors, and each model receives a single number that characterizes how well they predict the neural responses across all test images. Accordingly, Brain-Score does not test any hypothesis regarding how size, position, or orientation are encoded in DNNs or humans. The other four studies that test DNN-IT correspondences used similar datasets, and again, Brain-Score averaged across all test images when generating predictions.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

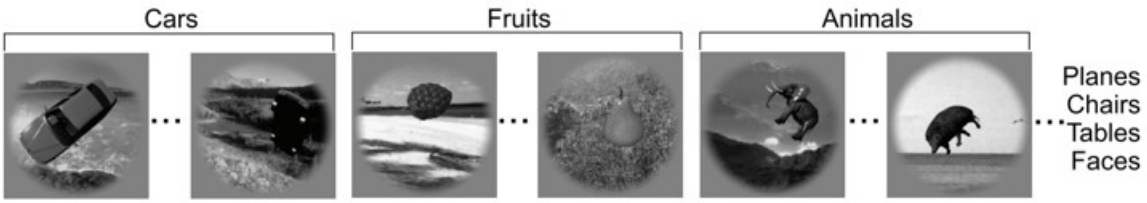


Figure 2. Example images of cars, fruits, and animals at various poses with random backgrounds from Majaj et al. (2015).

Similarly, consider the two studies in Brain-Score that assess how well DNNs predict behavior in humans and macaques. The first used objects displayed in various poses and randomly assigned backgrounds (similar to Figure 2; Rajalingham et al, 2015), but again, predictions were made after collapsing over the various poses. The second study was carried out by Geirhos et al., (2021) who systematically varied images across multiple conditions to test various hypotheses regarding how DNNs classify objects. For example, in one comparison, objects were presented as photographs or as line drawings to assess the role of shape in classifying objects (see Section 4.2.3). However, in Brain-Score, the performance of models is again averaged across all conditions such that the impact of specific manipulations is lost[1]. In sum, in all current prediction-based experiments, models are assessed in how well they predict overall performance, with the assumption that the higher the prediction the better the DNN-human (macaque) correspondence. On this approach, there is no attempt to assess the impact of any specific image manipulation, even when the original experiments specifically manipulated independent variables to test hypotheses.

This is not to say that researchers comparing DNNs to humans using prediction-based experiments do not manipulate any variables designed to test hypotheses. Indeed, the standard approach is to compare different DNNs on a given dataset; in this sense, the researcher is manipulating a theoretically motivated

[1] The Brain-Score website currently lists 18 behavioral benchmarks, but the data were taken from just the Rajalingham et al. and Geirhos et al. papers, with 17 image manipulations from the later study all described as separate benchmarks. However, it should be noted that the two papers each contributed equally to the overall behavioral benchmark score, with the mean results over the 17 conditions weighted equally with the Rajalingham et al. findings.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

variable (the models). However, these manipulations tend to compare models that vary along multiple dimensions (architectures, learning rules, objective functions, etc.) rather than assess the impact of a specific manipulation (e.g., the impact of pretraining on ImageNet). Accordingly, it is rarely possible to attribute any differences in predictivity to any specific manipulation of the models. And even when the modeler does run a controlled experiment in which two models are the same in all respects apart from one specific manipulation (e.g., Mehrer et al., 2020), the two models are still being assessed in a prediction-based experiment where there is no assessment of what visual properties of the images are driving the predictions.

This method of evaluating DNNs as models of human vision and object recognition is at odds with general scientific practice. Most research is characterized by running controlled experiments that vary independent variables to test specific hypotheses regarding the causal mechanisms that characterize some natural system (in this case, biological vision). Models are supported to the extent that they account for these experimental results, among other things. The best empirical evidence for a model is that it survives "severe" tests (Mayo, 2018), namely, experiments that have a high probability of falsifying a model if and only if the model is false in some relevant manner. Relatedly, models are also supported to the extent that they can account for a wide range of qualitatively different experimental results because there may be multiple different ways to account for one set of findings but far fewer ways to explain multiple findings. Of course, prediction is also central to evaluating models tested on controlled experiments, but prediction takes the form of accounting for the experimental results of studies that manipulate independent variables, with prediction in the service of explanation. That is, the goal of a model is to test hypotheses about how a natural system works rather than account for the maximum variance on behavioral and brain datasets.

Outside the current DNN modeling of human vision and object recognition there are few areas of science where models are assessed on prediction-based experiments and compete on benchmark datasets with the assumption that, all else being equal, models with better predictions more closely mirror the system under

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

investigation. There are fewer areas still where prediction-based experiments drive theoretical conclusions when it is possible to perform controlled experiments that vary independent variables designed to test specific hypotheses. Even the simpler Parallel Distributed Processing (PDP) network models developed in the 1980s were assessed on their ability to account for a wide range of experimental results reported in psychology (McClelland, Rumelhart, & PDP Research Group, 1986).

Our contention is that researchers should adopt standard scientific methods and assess models on their ability to accommodate the results of controlled experiments from psychology (and related disciplines) rather than on prediction-based experiments. We not only show that there are principled and practical problems with the current approach, but also, that many of the inferences drawn from prediction-based experiments are in fact wrong.

2.1 The "in principle" problems with relying on prediction when comparing humans to DNNs:

There are three fundamental limitations with prediction-based experiments that undermine the strong claims that are commonly made regarding the similarities between DNN and human object recognition.

2.1.1: Correlations do not support causal conclusions. Scientists are familiar with the phrase "correlation does not imply causation", but the implication for DNN modeling is underappreciated, namely, good predictions do not entail that two systems rely on similar mechanisms or representations (admittedly, not as snappy a phrase). Guest and Martin (2021) give the example of a digital clock predicting the time of a mechanical clock. One system can provide an excellent (in this case perfect) prediction of another system while relying on entirely different mechanisms. In the same way, DNN models of object recognition that make good (even perfect) predictions on behavioral and brain datasets might be poor models of vision. In the face of good predictions, controlled experiments that manipulate independent variables designed to test hypotheses are needed to determine whether the two systems share

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

similar mechanisms. In the current context, it is the most straightforward way to assess whether a DNN that tops the rankings on a benchmark like Brain-Score is computing in a brain-like manner.

How seriously should we take this objection? If something walks like a duck and quacks like a duck, isn't it in all likelihood a duck? In fact, DNNs often make their predictions in unexpected ways, exploiting "short-cuts" that humans never rely on (e.g., Geirhos et al., 2018; Malhotra, Evans, and Bowers, 2020; Malhotra, Dujmović, & Bowers, 2022; Rosenfeld, Zemel, Tsotsos, 2018). For example, Malhotra et al. (2020) systematically inserted single pixels (or clouds of pixels) into photographs in ways that correlated with image category so that the images could be classified based on the photographic subjects themselves or the pixels. DNNs learned to classify the images based on the pixels rather than the photos, even when a single pixel was nearly imperceptible to a human. In all cases of short-cuts, the performance of DNNs is mediated by processes and/or representations that are demonstrably different from those used in biological vision.

The critical issue for present purposes, however, is whether models that classify images based on short-cuts also perform well on prediction-based experiments. Dujmović et al. (2022) explored this question using RSA which compares the distances between activations in one system to the distances between corresponding activations in the second system (see Figure 3). To compute RSA, two different systems (e.g., DNNs and brains) are presented the same set of images and the distance between the representations for all pairs of images is computed. This results in two representational dissimilarity matrices (RDMs), one for each system. The similarity of these RDMs gives an RSA score. That is, rather than directly comparing patterns of activations in two systems, RSA is a second-order measure of similarity. In effect, RSA is a measure of representational geometry similarity – the similarity of relative representational distances of two systems. High RSA scores between DNNs and humans (and monkeys) have often been used to conclude that these systems classify images in similar ways (e.g., Kiat et al., 2022, Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi, & Kriegeskorte, 2014; Kriegeskorte et al., 2008).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

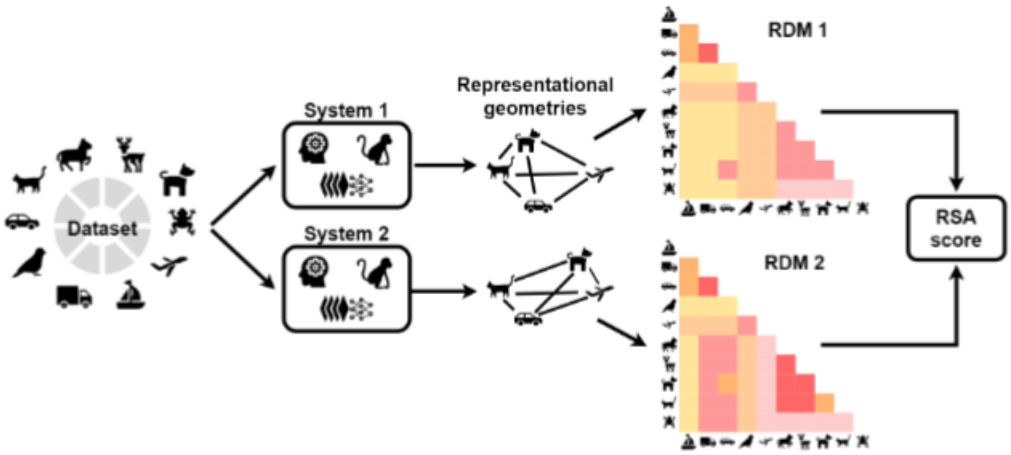


Figure 3. RSA calculation. A series of stimuli from a set of categories (or conditions) are used as inputs to two different systems (for example, a brain and a DNN). The corresponding neural or unit activity for each stimulus is recorded and pairwise distances in the activations within each system are 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 (image taken from Dujmović et al., 2022).

To assess whether large RSAs can be obtained between two very different systems, Dujmović et al. (2022) carried out a series of simulations that computed RSAs between two DNNs or between DNNs and single-cell recordings from macaque IT when the two systems classified objects in qualitatively different ways. For example, when comparing DNNs to macaque IT, the authors trained a DNN to classify photographs taken from Majaj, Hong, Solomon, and DiCarlo (2015) that contained a pixel patch confound (call it DNN-pixel) as well as unperturbed photos (DNN-standard), similar to the Malhotra et al. (2020) setup described above. The critical finding was that RSAs could be pushed up or down systematically depending on the pixel patch locations. For certain placements of the patches, the RSA observed between the DNN-pixel and macaque IT matched the RSA scores achieved by networks pre-trained on naturalistic stimuli (ImageNet dataset) and fine-tuned on the unperturbed images (Figure 4, left). That is, even macaque IT and DNNs that classified objects based on single pixel patches could share representational geometries (for related discussion, see Kriegeskorte & Wei, 2021; Palmer, 1999). By contrast, the location of the patches on the DNN-standard network did not impact on RSAs.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

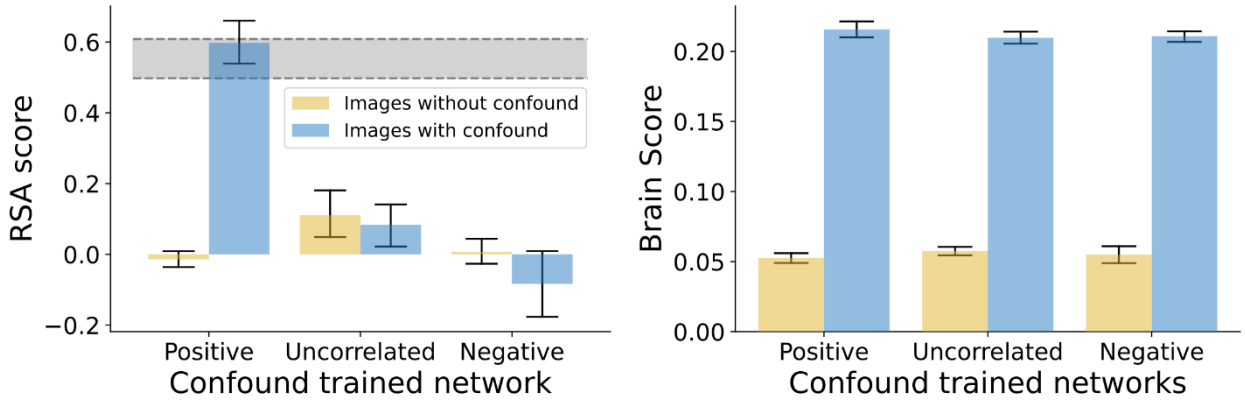


Figure 4. RSA (Left) and Brain-Score (Right) for networks trained on predictive pixels. The location of the pixel patches varied across conditions, such that the location was positively, negatively, or uncorrelated with the representational distances between classes in the macaque IT. When the pixel distances are positively correlated in the training set, RSA scores approached scores achieved by networks pre-trained on ImageNet and fine-tuned on unperturbed images. When the training images did not contain the pixel confounds, the location of the pixels at test did not impact on RSA scores. The dataset dependence of RSA scores extends to neural predictivity as measured by Brain-Score as the same pixel networks explain significantly more macaque IT activity when the confounding feature is present in the stimuli (RSA scores taken from Dujmović et al., 2022, Brain-Score results are part of ongoing, unpublished research).

Another common prediction method involves directly fitting unit activations from DNNs to brain activations (single-cell recordings or voxels in fMRI)in response to the same set of images using linear regression (e.g., Yamins et al., 2014). This neural predictivity approach is used in the Brain-Score benchmark (Schrimpf et al., 2020ab). Despite this important distinction between RSA and neural activity, when these two methods are used on behavioral and brain datasets they are both correlational measures, so again, it is possible that confounds are driving brain predictivity results as well. Consistent with this possibility, DNNs that classify images based on confounding features often perform well on Brain-Score. For example, object shape and texture are confounded in the natural world (and in ImageNet), with DNNs often classifying objects based on their texture and humans based on their shape (Geirhos et al. 2019; for more details see Section 4.1.2). Just as texture representations are used to accurately predict object categories in DNNs, texture representations in DNNs may be used to predict shape representations in the human (and macaque) visual system to obtain high neural predictivity scores. More direct evidence for

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

this comes from ongoing work by Dujmović et al. (unpublished) that has shown that neural predictivity is indeed influenced by confounding factors. For example, the ability of DNNs to predict macaque neural activity depended heavily on whether the images contained a confounding feature – in which case predictivity rose drastically compared to when the confound was not present (See Figure 4, right). In this case, the spatial organization of the confounding pixel patches did not matter, presumably reflecting the fact that neural predictivity does not assess the similarity representational geometries. Thus, a good neural predictivity score may reflect the fact that DNNs are exploiting confounds (short-cuts) in datasets rather than mirroring biological vision.

It is not only the presence of confounds that can lead to misleading conclusions based on predictions. Another factor that may contribute to the neural predictivity score is the effective latent dimensionality of DNNs – that is, the number of principal components needed to explain most of the variance in an internal representation of DNNs. Elmoznino & Bonner (2022) have shown that effective latent dimensionality of DNNs significantly correlates with the extent to which they predict evoked neural responses in both the macaque IT cortex and human visual cortex. Importantly, the authors controlled for other properties of DNNs, such as number of units in a layer, layer depth, pre-training, training paradigm, etc. and found that prediction of neural data increases with an increase in effective dimensionality, irrespective of any of these factors. In other words, DNNs may outperform other models on benchmarks such as Brain-Score not because their internal representations or information processing is similar to information processing in the cortex, but because they effectively represent input stimuli in higher dimensional latent spaces.

Of course, two DNNs (or a DNN and a brain) that do represent objects in a highly similar way will obtain high RSAs and high neural predictivity scores, but the common assumption that high RSAs and predictivity scores indicate that two systems work similarly is unsafe. This is illustrated in Figure 5 where better performance on prediction-based experiments can correspond to either more or less similarity to human vision, and where models with benchmark scores of zero can provide important insights into

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

human vision (because the model does not even take images as inputs). The most straightforward way to determine whether good performance on prediction-based experiments reflects meaningful DNN-brain correspondences is to carry out controlled experiments.

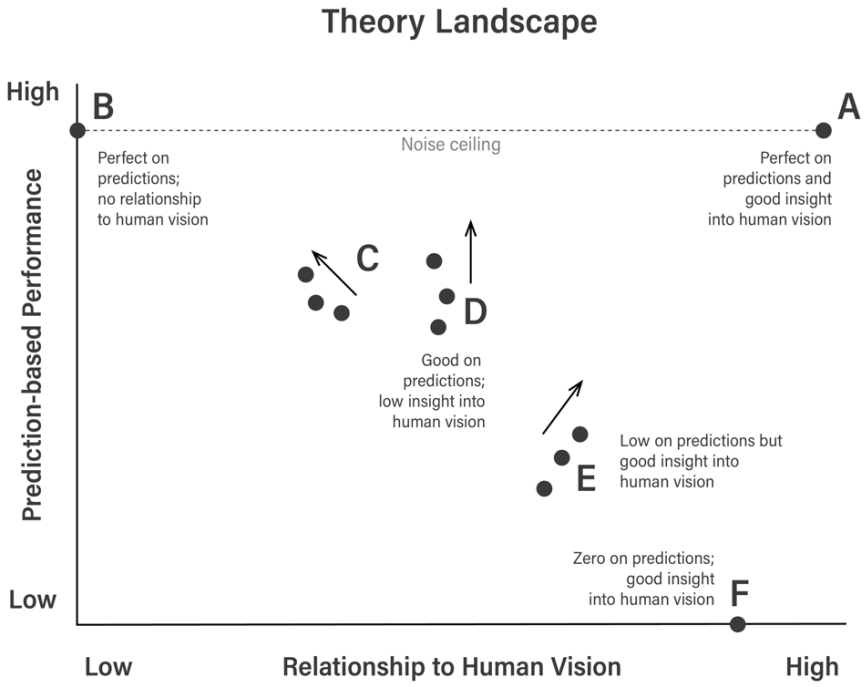


Figure 5. Different models fall in different parts of the theory landscape. Critically, it is possible to do well on prediction-based experiments despite poor correspondences to human vision, and there is no reason to expect that modifying a model to perform better on these experiments will necessarily result in better models of human vision. Similarly, poor performance does not preclude the model from sharing important similarities with human vision. Noise ceiling refers to how well humans predict one another on prediction-based experiments, and it is the best one can expect a model to perform.

2.1.2: Prediction-based experiments provide few theoretical insights: Putting aside the misleading estimates of DNN-human similarity that may follow from prediction-based experiments, the theoretical conclusions one can draw from good predictions are highly limited compared to cases in which models are tested against controlled experiments. For example, perhaps the most fundamental finding regarding human basic-level object recognition is that we largely rely on shape representations (Biederman & Ju, 1988). This results in humans recognizing objects based on their shape rather than texture when the texture of one category is superimposed on the shape of another (e.g., an image that takes the shape of a

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

cat and a texture of an elephant is classified as a cat; Geirhos et al., 2019; for more details see Section 4.1.2). Importantly, a model's success or failure in capturing this result is theoretically informative. In the case of a success, the model may provide some insight into how shape is encoded in the visual system. And when a model fails, it can provide guidance for future research (e.g., researchers can try to modify the training environments, architectures of DNNs, etc., in theoretically motivated ways to induce a shape bias).

By contrast, no similar insights derive from high scores on prediction-based experiments (even assuming the good predictions provide an accurate reflection of DNN-brain similarity). For example, it is not clear whether the models at the top of the Brain-Score leaderboard classify images based on shape or texture. To answer this question, some sort of controlled experiment needs to be carried out (such as the Geirhos et al., 2019 controlled experiment). More generally, when a DNN falls short of the noise ceiling on prediction-based experiments the limited success does not provide specific hypotheses about how to improve the model. Researchers might hypothesize that DNNs should be trained on more ecological datasets (e.g., Mehrer, 2021), or that it is important to add top-down connections that characterize the human visual system (e.g., Zhuang et al., 2021), etc. However, the size of the gap between performance and the noise ceiling does not suggest which of the different possible research directions should be pursued, or which of multiple different dimensions of variations between models (e.g., the architecture, learning rule, optimization function, etc.) is most responsible for the failure (or success).

2.1.3: Prediction-based experiments restrict the types of theories that can be considered: Finally, the reliance on current prediction-based experiments ensures that only "image computable" models that can take photorealistic images as inputs are considered. This helps explain why psychological models of object recognition are ignored in the DNN community. By contrast, when assessing models on their ability to account for results of controlled experiments, a broader range of models can be assessed and compared. For example, consider the *recognition by components* (RBC) model of basic-level object

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

recognition that was first formulated at a conceptual level to explain a wide variety of empirical findings (Biederman, 1987) and later elaborated and implemented in a neural network architecture called JIM (Hummel & Biederman, 1992). These two models could not be more different than current DNNs given that they characterize the representations, processes, and even objective functions in qualitatively different ways. Nevertheless, the RBC and JIM models make multiple predictions regarding human object recognition and vision more generally, and accordingly, can be compared to DNNs in terms of their ability to predict (and explain) a wide variety of empirical phenomena (of the sort reviewed in Section 4). The common conclusion that DNNs are the best models of human object recognition relies on excluding alternative models that do account for a range of key experimental results reported in psychology.

To summarize, the common claim that DNNs are currently the best models of human vision relies on prediction-based experiments that may provide misleading estimates of DNN-human similarity, that provide little theoretical insight into the similarities that are reported, and that exclude the consideration of alternative models that do explain some key empirical findings. It is important to emphasize that these principled problems do not only limit the conclusions we can draw regarding the current DNNs tested on prediction-based experiments and benchmarks such as Brain-Score (at the time of writing over 200 DNNs have been submitted to the Brain-Score leaderboard with models spanning a wide variety of architectures and objective functions). These problems will apply to any future model evaluated by prediction-based experiments.

2.2 The practical problems with prediction when comparing humans to CNNs

Apart from the principled problems of comparing DNNs to humans using current prediction-based experiments, there are also a variety of methodological issues that call into question the conclusions that are often drawn. With regards to prediction-based experiments on brain data, perhaps the most obvious practical problem is the relative scarceness of neural data on which the claims are made. For example, as

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

noted above, the Brain-Score match to high-level vision in IT is based on 5 studies that rely on a total of 3 monkeys presented with two very similar image datasets. Similarly, the reports of high RSAs between DNNs and human vision has largely relied on a small set of studies, and these studies often suffer methodological limitations (Xu & Vaziri-Pashkam, 2021). This raises the concern that impressive predictions may not generalize to other datasets, and indeed, there is some evidence for this. For example, Xu and Vaziri-Pashkam (2021) used a more powerful fMRI design to assess the RSA between DNNs and human fMRI for a new dataset of images, including images of both familiar and novel objects. They found the level of correspondence was much reduced compared to past studies. For familiar objects, they failed to replicate past reports that early layers of DNNs matched V1 processing best and later layers of DNNs matched later layers of visual cortex best. Instead, Xu and Vaziri-Pashkam only obtained high RSAs between early levels of DNNs and V1[2]. Similarly, with unfamiliar objects, Xu and Vaziri-Pashkam failed to obtain any high DNN-human RSA scores at any layers. These failures were obtained across a wide range of DNNs, including CORnet-S that has been described as the "current best model of the primate ventral visual stream" (Kubilius et al., 2019, p. 1) based on its Brain-Score. The impressive DNN-human RSAs reported in the literature may evidently not generalize broadly. For similar outcome in the behavioral domain see Erdogan and Jacobs (2017) discussed in section 4.1.9.

Another problem is that DNNs that vary substantially in their architectures support similar levels of predictions (Storrs et al., 2021). Indeed, even untrained networks (models that cannot identify any images) often support relatively good predictions on these datasets (Truzzi, & Cusack, 2020), and this may simply reflect the fact that good predictions can be made from many predictors regardless of the similarity of DNNs and brains (Elmoznino & Bonner, 2022). Furthermore, when rank ordering models in terms of their (often similar) predictions, different outcomes are obtained with different datasets. For example, there is

[2] Interesting, some classical models of V1 processing do substantially better in accounting for the V1 responses compared to DNNs when assessed on the Brain-Score dataset itself. Go to: http://www.brain-score.org/competition/#workshop

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

only a .42 correlation between the two V1 benchmark studies listed on the current Brain-Score leaderboard. Consider just one network: mobilenet_v2_0.75_192 achieves a neural predictivity score of .783 on one V1 dataset (ranking in the top 10) and .245 on another (outside the top 110). Given the contrasting rankings, it is not sensible to conclude that one model does a better job in predicting V1 activity by simply averaging across only two benchmarks, and more generally, these considerations highlight the problem of ranking networks based on different scores.

In addition, there are issues with the prediction-based experiments carried out on behavioral studies showing that DNNs and humans make similar classification errors (e.g., Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016; Kubilius, Bracci, & Op de Beeck, 2016; Rajalingham, Schmidt, & DiCarlo, 2015; Rajalingham et al.2018; Tuli, Dasgupta, Grant, & Griffiths, 2021). Geirhos, Meding, and Wichmann (2020) argue that the standard methods used to assess behavioral correspondences have led to inflated estimates, and to address this concern, they adapted an error consistency measure taken from psychology and medicine where inter-rater agreement is measured by Cohen's kappa (Cohen, 1960). Strikingly, they reported near chance trial-by-trial error consistency between humans and a range of DNNs. This was the case even with CORnet-S that has one of the highest overall behavioral Brain-Scores. More recently, error consistency was found to improve in DNNs trained on much larger datasets, such as CLIP that is trained on 400 million images (Geirhos et al., 2021). Nevertheless, the gap between humans and the best performing DNN was substantial. For example, the if you consider the top-10 performing models on the Brain-Score leaderboard, the error consistency between DNNs and humans for edge filtered images (images that keep the edges but remove the texture of images) is .17. Clearly, the different methods used to measure behavioral consistency provide very different conclusions, and the DNN-human correspondences for some types of images that humans can readily identify remain very low.

3 The theoretical problem with DNNs as models of human object recognition

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

Apart from the principled and practical problems with prediction-based experiments, the general approach of modelling human object recognition by optimizing classification performance may be misguided for a theoretical reason, namely, the human visual system may not be optimized to classify images. For example, Malhotra et al. (2021) argue that the human visual system is unconcerned with the proximal stimulus (the retinal image) except inasmuch as it can be used to make inferences about the distal stimulus (the object in the world) that gave rise to it. The advantage of distal representations is that they afford a wide range of capacities beyond image classification, including visual reasoning (e.g., Hummel, 2013). The downside is that constructing distal representations is an ill-posed problem, meaning it cannot be solved based on the statistics available in the proximal stimuli alone, or in the mapping between the proximal stimulus and, say, an object label. Accordingly, on this view, the visual system relies on various heuristics to estimate the distal properties of objects, and these heuristics reveal themselves in various ways, including Gestalt rules of perceptual organization (see Section 4.2.3) and shape processing biases (see Section 4.1.4). It is unclear whether the relevant heuristics can be learned by optimizing classification performance, and at any rate, current DNNs do not acquire these heuristics, as discussed below.

Furthermore, even if building distal representations from heuristics is a misguided approach to understanding human object recognition, it is far from clear that optimizing on classification is the right approach. Indeed, evolution (which may be considered as an optimization process) rarely (if ever) produces a cognitive or perceptual system in response to a single selection pressure. Rather, evolution is characterized by "descent through modification" with different selection pressures operating at different times in our evolutionary history (Marcus, 2009; Zador, 2019). This results in solutions to complex problems that would never be found if a single selection process was operative from the start. Marcus (2009) gives the example of the human injury-prone spinal column that was a modification of a horizontal spine designed for animals with four legs. Better solutions for bipedal walkers can be envisaged, but the

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

human solution was constrained by our ancestors. See Marcus (2009) for a description of the many foibles of the human mind that he attributes to a brain designed through descent with modification.

Furthermore, evolutionary algorithms can produce solutions to complex problems when there is no selection pressure to solve the problem at all. For example, Lehman and Stanley (2011) used evolutionary algorithms to produce virtual robots that walked. In one condition the selection pressure was to walk as far as possible and in another the selection pressure was behavioral "novelty", that is, robots that did something different from all other robots. Despite the lack of any selection pressure to walk, the latter robots walked further. Lehman and Stanley (2011) reported similar outcomes in other domains such as solving mazes, with virtual robots selected to produce novel behaviors doing much better than models selected to solve mazes. Moreover, compared to selecting for the desired outcome directly, novelty search evolved more complex and qualitatively different representations (Woolley & Stanley, 2011). The explanation for these counter-intuitive findings is that the search environment is often "deceptive", meaning that optimizing on the ultimate objective will often lead to dead ends. In some cases, the only way to find a solution to an objective (e.g., walking) is to first evolve an archive of architectures and representations that may all appear irrelevant to solving the objective (so-called "stepping stones"; Stanley et al., 2019), and it may require different selection pressure(s) than optimizing for the objective itself.

Even though the human visual system is the product of multiple selection pressures, all the top-performing models on Brain-Score and related prediction-based experiments were just optimized to classify objects. Of course, these DNNs do have "innate" structures generally composed of a collection of convolution and pooling operators, but these structures are largely chosen because they improve object recognition on ImageNet and other image datasets. Furthermore, despite the fact that convolutions and pooling are loosely inspired by neuroscience, the architectures of DNNs are radically different from brain structures in countless ways (Izhikevich 2004), 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

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

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. This is even more so for recent state-of-the-art Transformer models of object recognition (Tuli et al., 2021) that do not even include innate convolution and pooling operators.

It is not a safe assumption that these (and countless other) different starting points do not matter, and that optimizing on classification will bridge the difference between DNNs and human object recognition. Similarly, more recent self-supervised networks are first optimized to predict their visual inputs and only subsequently optimized to classify the images, but again, it is far from clear that self-supervision provides the right starting point to optimize on classification. A related critique has been applied to Bayesian theories in psychology and neuroscience according to which minds and brains are (near) optimal in solving a wide range of tasks. Again, little consideration is given to descent with modification or physiological constraints on solutions, and this can lead to "just so" stories where models account for human performance on a set of tasks despite functioning in qualitatively different ways (Bowers & Davis, 2012ab; for response see Griffiths, Chater, Norris, & Pouget, 2012).

This theoretical concern should be considered in the context of the principled and practical problems of evaluating models on prediction-based experiments on behavioral and brain studies. That is, not only is it possible that DNNs and humans identify objects in qualitatively different ways despite good predictions, but there are also good reasons to expect that they do. As we show next, the empirical evidence strongly suggests that current DNNs and humans do indeed identify objects in very different ways.

4 The empirical problem with claiming DNNs and human vision are similar

These principled, practical, and theoretical issues do not rule out the possibility that current DNNs are good or even the current best models of human vision and object recognition. Rather, they imply that the

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

evidence from this approach is ambiguous and strong conclusions are not yet justified. What is needed are controlled experiments to better characterize the mechanisms that support DNN and human object recognition.

In fact, some researchers have assessed how well models account for the results of controlled experiments in psychology in which images have been manipulated to test specific hypotheses. In some cases the behavior of a model (that is, the model's output) is compared with human behavior, and in other cases, the activations of hidden units within a model are compared to perceptual phenomena reported by humans. Although these findings are largely ignored by modelers focused on brain-prediction studies, it is striking how often these studies highlight stark discrepancies between DNNs and humans, and how informative these studies are for developing better models of human vision. In this section we review multiple examples of DNNs failing to account for key experimental results reported in psychology. We also review key psychological phenomena that have largely been ignored and that require more investigation.

4.1 Discrepancies:

4.1.1 DNNs are highly susceptible to Adversarial attacks: Adversarial images provide a dramatic example of an experimental manipulation that reveals a profound difference between human and DNN object recognition. Adversarial images can be generated to look unfamiliar to humans but that nevertheless fool DNNs into confidently classifying them as members of familiar categories (See Figure 6). These images do not appear in behavioral benchmarks such as those used in Brain-Score, and if they were, they would undermine any claim that humans and DNNs make similar errors when classifying images. Some researchers have pointed out that humans experience visual illusions, and adversarial attacks might just be considered a form of illusion experienced by DNNs (Kriegeskorte, 2015). However, these "illusions" are nothing like the illusions experienced by humans. Although there have been some reports that humans and DNNs encode adversarial images in a similar way (Zhou & Firestone, 2019),

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

careful behavioral studies show this is not the case (Dujmović, Malhotra, & Bowers, 2020). There has been some limited success at making DNNs more robust to adversarial attacks by explicitly training models to not classify these images as familiar categories. But it is not necessary to train humans in this way. What is needed is a psychologically plausible account that fully addresses the problem.

Figure 6. Example of adversarial images for three different stimuli generated in different ways. In all cases the model is over 99% confident in its classification. Images taken from Nguyen, Yosinski, and Clune (2015).

4.1.2 DNNs often classify images based on texture rather than shape: A fundamental conclusion from psychological research is that humans largely rely on shape when identifying objects. Indeed, adults classify line drawings of objects as quickly as colored photographs (Biederman & Ju, 1988), and infants can recognize line drawings the first time they are seen (Hochberg & Brooks, 1962). Accordingly, a model of human object recognition should largely rely on shape when classifying objects. However, this is not the case for most DNN models that perform well on Brain-Score and other prediction metrics. For example, Geirhos et al. (2019) developed a "style transfer" dataset where the textures of images from one category were superimposed on the shapes of images from other categories (e.g., a shape of a cat with the texture of an elephant) to assess the relative importance of texture vs shape on object recognition. Unlike humans, DNNs trained on natural images relied more on texture (e.g., classifying a cat-elephant image as an elephant; See Figure 7). Indeed, the CORnet-S model described as one of the best models of human vision largely classifies objects based on texture (Geirhos, Meding, & Wichmann, 2020), and this contrast

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

between DNNs and humans extends to children and adults (Huber, Geirhos, & Wichmann 2022; but see Ritter et al., 2017, for the claim that DNN have a human-like shape-bias).

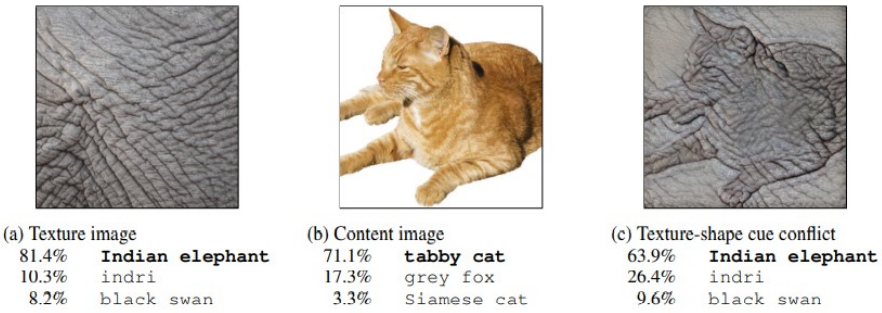


Figure 7. Illustration of a style-transfer image in which (a) the texture of an elephant and (b) the shape of a cat that combine to form (c) the shape of a cat with the texture of an elephant. The top three classifications of a DNN to the three images are listed below each image, with the model classifying the style-transfer image as an elephant with 63.9% confidence (the cat is not in the top three choices of the DNN that together account for 99.9% of its confidence). Images taken from Geirhos et al., 2019.

More recently, Malhotra et al. (2022) compared how DNNs and humans learn to classify a set of novel stimuli defined by shape as well as one other non-shape diagnostic feature (including patch location and segment color as shown in Figure 8). Humans showed a strong shape-bias when classifying these images, and indeed, could not learn to classify the objects based on some non-shape features. By contrast, DNNs had a strong bias to rely on these very same non-shape features. Importantly, when the DNNs were pre-trained to have a shape bias (by learning to classify a set of images in which shape but not texture was diagnostic of object category), the models nevertheless focused on non-shape features when subsequently trained to classify these stimuli. This was the case even after freezing the convolutional layers of a shape biased ResNet50 (that is, freezing 49 of the 50 layers of the DNN). This suggests that the contrasting shape biases of DNNs and humans is not the product of their different training histories as sometimes claimed (Hermann, Chen, & Kornblith, 2019).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

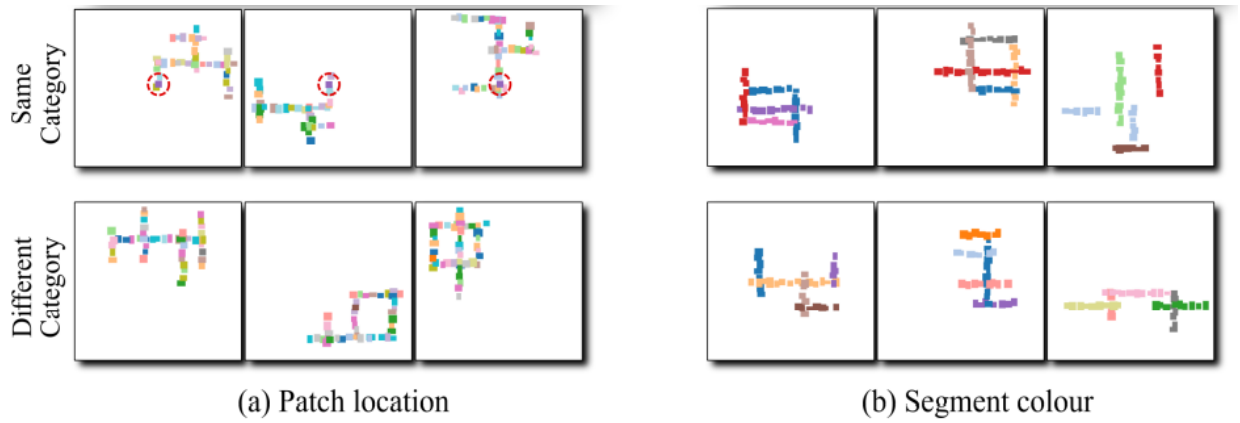
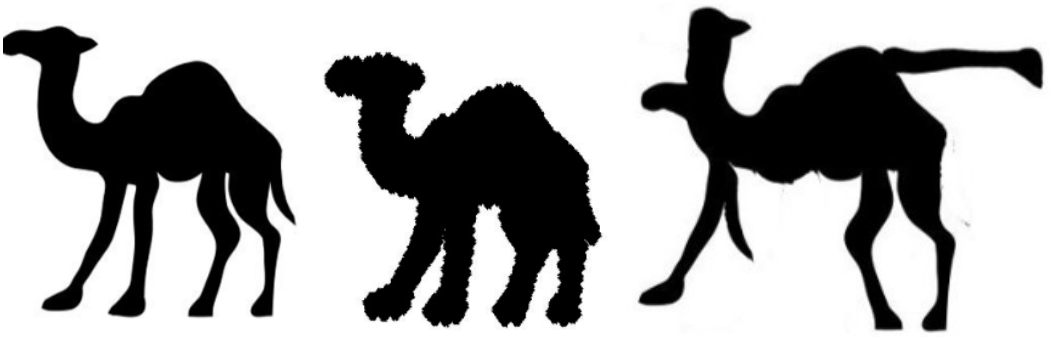


Figure 8. Examples of novel stimuli defined by shape as well as one other non-shape feature. In (a) global shape and location of one of the patches define a category, and for illustration, the predictive patch is circled. Stimuli in the same category (top row) have a patch with the same color and the same location, while none of the stimuli in any other category (bottom row) have a patch at this location. In (b) global shape and color of one of the segments predicts stimulus category. Only stimuli in the same category (top row) but not in any other category (bottom row) have a segment of this color (red). The right-most stimulus in the top row shows an example of an image containing a non-shape feature (red segment) but no shape feature. Images taken from Malhotra et al., 2022.

4.1.3 DNNs classify images based on local rather than global shape: Although DNNs rely more on texture than shape when classifying naturalistic images (images in which both shape and texture are diagnostic of category), several studies have shown that modifying the learning environment (Geirhos et al., 2019; Hermann et al., 2019) or architecture (Evans et al., 2021) of DNNs can increase the role of shape in classifying naturalistic images. Nevertheless, when DNNs classify objects based on shape, they use the wrong sort of shape representations. For instance, in contrast with a large body of research showing that humans tend to rely on the global shape of objects, Baker, Lu, Erlikhman, and Kellman (2018) showed that DNNs focus on local shape features. That is, they found that DNNs trained on ImageNet could correctly classify some silhouette images (where all diagnostic texture information was removed), indicating that these images were identified based on shape. However, when the local shape features of the silhouettes were disrupted by including jittered contours, the models did much more poorly. By contrast, DNNs were more successful when the parts of the silhouettes were rearranged, a manipulation that kept many local shape features but disrupted the overall shape. Humans show the opposite pattern. See Figure 9.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press



a) b) c)

Figure 9. Illustration of (a) a silhouette image of a camel, (b) and image of a camel in which local shape features were removed by including jittered contours, and (c) and image of a camel in which global shape was disrupted. The DNNs had more difficulty in conditions (b) than (c). Images taken from Baker et al. (2018).

4.1.4 DNNs ignore the relations between parts when classifying images: Another key property of human shape representations is that the relations between object parts play a key role in object recognition. For example, Hummel and Stankiewicz (1996) trained participants to identify a set of "Basis" objects that were defined by their parts and the relation between the parts, and then assessed generalization on two sets of images: (1) Relational variants that were highly similar in terms of pixel overlap but differed in a categorical relation between two parts, and (2) Pixel variants that differed more in terms of their pixel overlap but shared the same set of categorical relations (see Figure 10). Across five experiments participants frequently mistook the Pixel variants as the Basis objects but rarely the Relational variants, indicating that the human visual system is highly sensitive to the relations. By contrast, when DNNs were trained on the Basis objects, the models mistook both the Relational and Pixel variants as the Basis objects and were insensitive to the relations (Malhotra et al., 2021). This was the case even after explicitly training the DNNs on these sorts of relations. As noted by Malhotra et al., the human encoding of relations between object parts may be difficult to achieve with current DNNs and additional mechanisms may be required.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

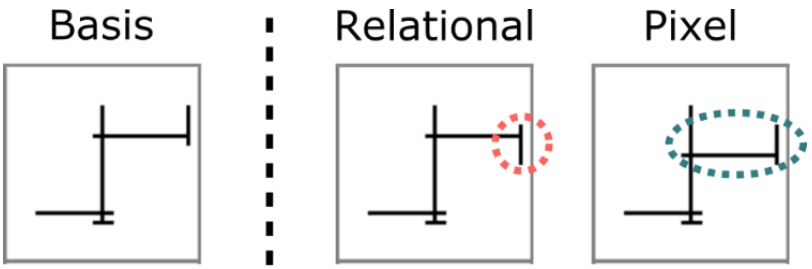


Figure 10. An example of (a) a Basis object, (b) a Relational Variant object that was identical to the Basis object except that one line was moved so that its "above/below" relation to the line to which it was connected changed (from above to below or vice-versa), as highlighted by the circle, and (c) a Pixel Variant object that was identical to the Basis object except two lines were moved in a way that preserved the categorical spatial relations between all the lines composing the object, but changed the coordinates of two lines, as highlighted by the oval. Images taken from Malhotra et al. (2021).

4.1.5 DNNs fail to distinguish between boundaries and surfaces: In human vision boundaries and surfaces of objects are processed separately and then combined early in the visual processing stream to perceive colored and textured objects. This separation is observed in V1 with neurons in the "interblobs" system coding for line orientations independent of color and contrast and neurons in a "blob" system coding for color in a way that is less dependent on orientation (Livingston and Hubel, 1988). A wide variety of color, lightness, and shape illusions are the product of the interactions between these two systems (Grossberg & Mingolla, 1985), with no explanation offered in DNNs that fail to factorize shape and color in two parallel streams. See Figure 11 for a striking example of surface filling in from boundaries. Importantly, filling-in occurs early, such that illusory surfaces can "pop-out", a signature that the process occurs before an attentional bottleneck constrains parallel visual processing (Ramachandran, 1992). The entanglement of shape and color representations in CNNs may also help explain why DNNs do not have a strong shape bias when classifying objects.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

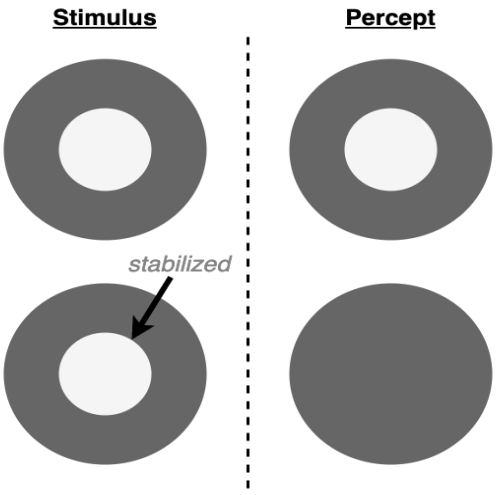


Figure 11. The phenomenon of filling-in suggests that edges and textures are initially processed separately and then combined to produce percepts. In this classic example from Krauskopf (1963), an inner green disc (depicted in white) is surrounded by a red annulus (depicted in dark grey). Under normal viewing conditions the stimulus at the top left leads to the percept at the top right. However, when the red-green boundary was stabilized on the retina as depicted in the figure in the lower left, subjects reported that the central disk disappeared and the whole target – disk and annulus – appeared red, as in lower right. That is, not only does the stabilized image (the green-red boundary) disappear (due to photo-receptor fatigue), but the texture from the outer annulus fills-in the entire surface as there is no longer a boundary to block the filling-in process. For more details see Pessoa, Tompson, and Noe (1998).

4.1.6 DNNs fail to show uncrowding: Our ability to perceive and identify objects is impaired by the presence of nearby objects and shapes, a phenomenon called crowding. For instance, it is much easier to identify the letter X in peripheral vision if it is presented in isolation compared to when it is surrounded by other letters, even if one knows where the letter is located. A more surprising finding is uncrowding, where the addition of more surrounding objects makes the identification of the target easier. Consider Figure 12 where participants are asked to perform a vernier discrimination task by deciding whether the top vertical line from a pair of vertical lines is shifted to the left or right. Performance is impaired when these lines are surrounded by a square rather than presented by themselves, an example of crowding. However, performance is substantially improved by the inclusion of additional squares, highlighting the role of long-range Gestalt like processes in which the squares are grouped together and then processed separately from the Vernier (Saarela, Sayim, Westheimer, & Herzog, 2009). Standard DNNs are unable to

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

explain uncrowding, but the LAMINART model of Grossberg and colleagues (e.g., Raizada, & Grossberg, 2001) designed to support grouping processes can capture some aspects of uncrowding (Francis, Manassi, & Herzog, 2017). Like the failure of DNNs to capture global shape, DNNs do not appear to encode the global organization of objects in a scene.

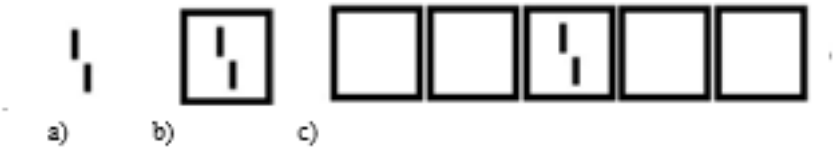


Figure 12. (a) In the standard vernier discrimination conditions two vertical lines are offset, and the task of the participant is to judge whether the top line is to the left or right of the bottom line. (b) In the crowding condition the vernier stimulus is surrounded by a square and discriminations are much worse. (c) In the uncrowding condition a series of additional squares are presented. Performance is much better here, although not as good as in (a).

4.1.7 DNNs are poor at identifying degraded and deformed images: Humans can identify objects that are highly distorted or highly degraded. For instance, we can readily identify images of faces that are stretched by a factor of four (Hacker & Biederman, 2018), when images are partly occluded or presented in novel poses (Biederman, 1987), and when various sorts of visual noise are added to the image (Geirhos et al., 2021). By contrast, CNNs are much worse at generalizing under these conditions (Alcorn et al., 2019; Geirhos et al., 2018, 2021; Wang et al., 2018; Zhu, Tang, Park, Park, & Yuille, 2019). It should be noted that more the largest DNNs do better on degraded images (e.g., CLIP trained on 400M images), but the types of errors the models make are still very different than humans (Geirhos et al., 2021).

4.1.8 DNNs have a superhuman capacity to classify unstructured data: While CNNs are too sensitive to various perturbations to objects, CNNs can learn to classify noise-like patterns at a super-human level. For example, Zhang, Bengio, Hardt, Recht, & Vinyals (2017) trained standard DNNs with ~1 million images composed of random pixel activations (TV static like images) that were randomly assigned to

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

1,000 categories. This shows that DNNs have a much greater capacity to memorize random data compared to humans, and this excess capacity may be exploited by DNNs to identify naturalistic images.

Tsvetkov, Malhotra, Evans, and Bowers (2020, in press) reduced the memorization capacities of DNNs by adding noise to the activation function (mirroring noise in neural activation), a bottleneck after the input canvas (analogous to the optic nerve where there are approximately 100 times fewer ganglion cells compared to photoreceptors), and using sigmoidal units that bound activation rather than rectified linear units common in state-of-the-art DNNs that can take on unbounded activation values. These modifications resulted in DNNs that were much better at learning to classify images from the CIFAR10 dataset compared to learning to classify random noise, consistent with human performance. At the same time, these networks were no better at classifying degraded CIFAR10 images. One challenge going forward will be to design DNNs that fail to learn random data but can identify degraded and deformed naturalistic images.

4.1.9 DNNs do not account for human similarity judgements for novel 3D shapes: There are various reports that DNNs provide a good account of human similarity judgments for familiar categories (Peterson, Abbott, & Griffiths, 2018; but see Geirhos et al., 2020). However, similarity judgements break down for unfamiliar objects. For example, German et al. (2020) measured human similarity judgements between pairs of novel part-based naturalistic objects (Fribbles) presented across multiple viewpoints. These judgments were then compared with the similarities observed in DNNs in response to the same stimuli. Overall, the degree of DNN-human similarity was only slightly better than would be predicted from a pixel-based similarity score, with accuracy near chance (under 58% with a baseline of 50%). Similar results were obtained by Erdogan and Jacobs (2017) when they assessed DNN-human similarity to novel 3D, cuboidal objects. The best similarity score was somewhat higher (64% with a baseline of 50%) and better than pixel-based similarity score, but much lower than an alternative Bayesian model which reached an accuracy of 87%. This no doubt relates to the observation that DNNs do not represent the

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

relations between object parts (Malhotra et al., 2021), a likely factor in the human similarity judgements for these multi-part 3D unfamiliar stimuli. Note, these behavioral outcomes are in line with the Xu and Vaziri-Pashkam (2021) results described above where they found that RSA scores between DNNs and fMRI signals were especially poor for unfamiliar objects.

4.1.10 DNNs fail to detect objects in a human-like way: Humans and CNNs not only classify objects but can also detect (and locate) objects in a scene. In the case of humans, there was an early report that object detection and object recognition occur at the same processing step in the visual system with Grill-Spector and Kanwisher (2005) concluding "as soon as you know it is there, you know what it is". Subsequent research addressed some methodological issues with this study and showed that humans can detect an object before they know what it is (Bowers & Jones, 2007; Mack, Gauthier, Sadr, & Palmeri, 2008). With regards to DNNs, there are multiple different methods of object detection, but in all cases we are aware of, detection depends on first classifying objects (e.g., Redmon, Divvala, Girshick, & Farhadi, 2016; Zou, Shi, Guo, & Ye, 2019). Why the difference? In the case of humans there are various low-level mechanisms that organize a visual scene prior to recognizing objects: Edges are assigned to figure or ground (Driver & Baylis, 1996), depth segregation is computed (Nakayama, Shimojo, & Silverman, 1989), nonaccidental properties such as colinearity, curvature, and cotermination, etc. are used to compute object parts (Biederman, 1987). These processes precede and play a causal role in object recognition, and these earlier processes presumably support object detection (explaining why detection is faster). The fact that CNNs recognize objects before detecting them suggests that they are lacking these earlier processes so central to human vision.

4.1.11 DNNs fail in same/different reasoning: The human visual system not only supports object recognition, but also visual reasoning (Hummel, 2000). Perhaps the simplest visual reasoning task is deciding whether two images are the same or different. Although there have been some recent reports that DNNs can support same/different judgements (Funke et al., 2021; Messina et al., 2021) the models were

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

only tested on images that were very similar to the training set. Puebla and Bowers (2022) provided a stronger test of whether DNNs support human-like same/different reasoning by testing DNNs on stimuli that differed from the training set (see Figure 13 for examples of images). The models failed when they were trained on stimuli taken from the set illustrated in the left-most panel of Figure 13 and tested on most other sets. Indeed, models failed on some test sets when trained to perform same/different judgements on stimuli from all sets but the test set. Even a network specifically designed to support visual relational reasoning, namely a Relation Network (Santoro et al., 2017), failed on some stimulus sets when trained on all others. For humans this is trivial without any training on the same/different task for any stimulus set.

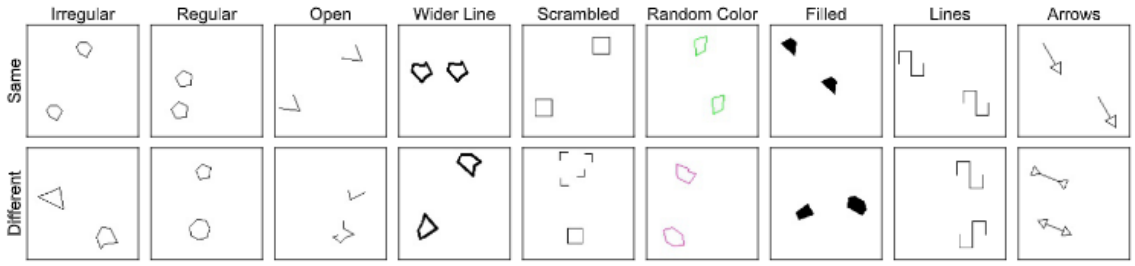


Figure 13. Example stimuli taken from 9 different stimulus sets, with Same trials depicted on the top row, Different trials on the bottom. The level of similarity between stimulus sets varied, with the greatest overlap between the Irregular and Regular sets, and little overlap between the Irregular set on the one hand and the Lines or Arrow datasets on the other. Image taken from Puebla and Bowers (2022).

4.1.12 DNNs are poor at visual combinatorial generalization: There are various reports that DNNs can support combinatorial generalization, but performance breaks down when more challenging conditions are tested. For example, Montero et al. (2021) explored whether DNNs that learn (or are given) "disentangled" representations (units that selectively encode one dimension of variation in a dataset) support the forms of combinatorial generalization that are trivial for humans. Despite the claim that disentangled representations support better combinatorial generalization (e.g., Duan et al., 2019), Montero et al. found a range of variational autoencoders trained to reproduce images succeeded in the simplest conditions but failed in more challenging ones. Indeed, DNNs with disentangled representations were no better than models using entangled (or distributed) representations. For example, after training to reproduce images of shapes on all locations except for squares on the right side of the canvas, the models

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

were unable to do so at test time, even though they had observed squares at other positions and other shapes at the right side. These results were consistent across other factor combinations and datasets and have been replicated using other training mechanisms and models (Schott et al, 2021). More recently, Montero et al. (2022) has shown that both the encoder and decoder components of variational autoencoders fail to support combinatorial generalization, and in addition, provide evidence that past reports of successes were in fact not examples of combinatorial generalization. There are still other models that appear to support combinatorial generalization in related conditions (Burgess et al., 2019; Greff et al., 2019), and it will be interesting to test these models under the conditions that disentangled models failed.

This pattern of success on easier forms of combinatorial generalization but failure on more challenging forms is common. For example, Barrett et al. (2018) assessed the capacity of various networks to perform Raven-style Progressive Matrices, a well-known test of human visual reasoning. Although the model did well in some conditions, the authors noted that a variety of state-of the art models (including Relational Networks designed to perform well in combinatorial generalization) did "strikingly poorly" when more challenging forms of combinatorial generalization were required. As noted by Greff et al. (2020), combinatorial generalization may require networks that implement symbolic processes through dynamic binding (currently lacking in DNNs) and they emphasize that better benchmarks are required to rule out any forms of short-cuts that DNNs might exploit (also see Montero et al., 2022, who identify conditions in which models appear to solve combinatorial tasks but fail when tested appropriately).

4.1.13 Additional failures on object recognition tasks: Perhaps the most systematic attempt to date to compare DNNs to psychological phenomena was carried out by Jacob, Pramod, Katti, & Arun (2020). They reported some correspondences between humans and DNNs (described in Section 4.28), but also a series of striking discrepancies. Amongst the failures, they showed DNNs trained on ImageNet do not encode the 3D shape of objects, do not represent occlusion or depth, and do not encode the part structure

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

of objects. For example, to investigate the representations of 3D shape, the authors presented pairs of images such as those in Figure 14 to DNNs. Humans find it easier to distinguish between the pair of images at the top of the figure compared to the pairs at the bottom even though each pair is distinguished by the same feature difference. The explanation is that humans perceive the former pair as 3D that take on different orientations whereas the latter stimuli are perceived as 2D. By contrast, DNNs do not represent the former pair as more dissimilar, suggesting that the models did not pick up on the 3D structure of these stimuli. Relatedly, Heinke, Wachman, van Zoest, and Leek (2021) showed that DNNs are poor at distinguishing between possible and impossible 3D objects, again suggesting DNNs fail to encode 3D object shape geometry.

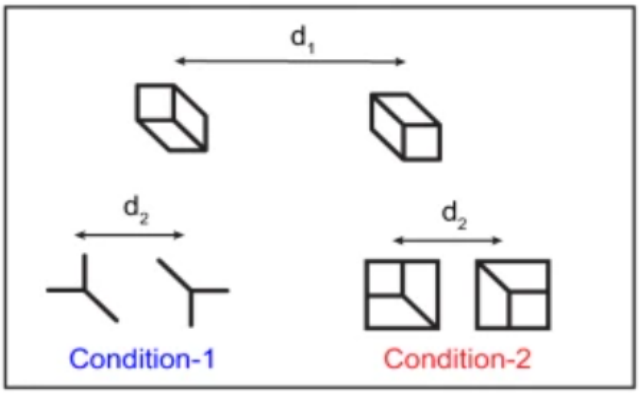


Figure 14. For humans the perceptual distance between the top pair of figures (marked as d1) is larger than the perceptual distance between the two pairs of objects on the bottom (marked as d2). For DNNs, the perceptual distance is the same for all pairs. Images taken from Jacob et al. (2020).

4.2 Key experimental phenomena that require more study before any conclusions can be drawn:

There are also a wide range of important psychological findings in vision that have received little consideration when assessing the similarities between human vision and DNNs. In a few of these cases there is some evidence that DNNs behave like humans, but the results remain preliminary and require more study before any strong conclusions are warranted. Here we briefly review some phenomena that should be further explored.

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

4.2.1 Perceptual Constancies: Human vision supports a wide range of visual constancies, including color, shape, and lightness constancies, where perceptual judgements remain stable despite changes in retinal input. For example, we often perceive the color of an object as stable despite dramatic changes in lighting conditions that change the wavelengths of light projected onto the retina. Similarly, we tend to perceive the size of an object as stable despite radical changes in the size of the retinal image when the object is viewed from nearby or far away. Perceptual constancies are critical to the visual system's ability to transform a proximal image projected on the retina into a representation of the distal object. Various forms of perceptual learning appear to operate on constancy-based perceptual representations rather than early sensory representations (Garrigan & Kellman, 2008). By contrast, it is not clear to what extent DNNs support perceptual constancies. Current evidence suggests that they do not, given that DNNs tend to learn the simplest regularities present in the input (e.g., Malhotra et al., 2021; Shah et al., 2020), and consequently, often learn short-cuts (Geirhos et al., 2020).

4.2.2 On-line Invariances: Human vision supports various visual invariances such that familiar objects can be identified when presented at novel scales, translations and rotations in the image plane, as well as rotations in depth. Furthermore, these invariances extend to untrained novel objects – what is sometimes called "on-line" invariance or tolerance (Blything, Biscione, & Bowers, 2020; Bowers, Vankov, & Ludwig, 2016). Although DNNs can be trained (Biscione & Bowers, 2021ab; Blything, Biscione, Vankov, Ludwig, & Bowers, 2020) or their architectures modified (Zhang, 2019) to support a range of on-line invariances, there are no experiments to date that test whether these models support invariance in a human-like way.

4.2.3 Gestalt principles: A wide range of Gestalt rules play a central role in organizing information in visual scenes, including organization by proximity, similarity, continuity, connectedness, and closure. That is, we do not just see the elements of a scene, we perceive patterns or configurations amongst the

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

elements, such that "the whole is more than the sum of its parts". This is not unique to the human cognitive architecture as some non-human animals show Gestalt effects (Pepperberg & Nakayama, 2016). Gestalt rules are not just some curiosity, they play a fundamental role in how we recognize objects by organizing the components of a scene (Biederman, 1987; Palmer, 2003; Wagemans et al, 2012). There are a few reports that DNNs are sensitive to closure (Kim, Reif, Wattenberg, & Bengio, 2021), although local features may mediate these effects (Baker, Kellman, Erlikhman, & Lu, 2018; Pang, O'May, Choksi, & VanRullen, 2021), and these effects only occur in the later layers of the network (whereas Gestalt closure effects can be detected in early human vision; Alexander, & Van Leeuwen, 2010). Biscione and Bowers (2022) provided some additional evidence that DNNs trained on ImageNet are indeed (somewhat) sensitive to closure in their later layers, but these same networks failed to support the Gestalt effects of orientation, proximity, and linearity, as illustrated in Figure 15. More work is needed to characterize which (if any) Gestalt effects are manifest in current DNNs. It is possible that differences in perceptual grouping processes may play a role in several additional DNN-human discrepancies, such as the failure of DNNs to identify objects based on global features, the failure of DNNs to show uncrowding, or the fact that DNNs classify objects before they detect them.

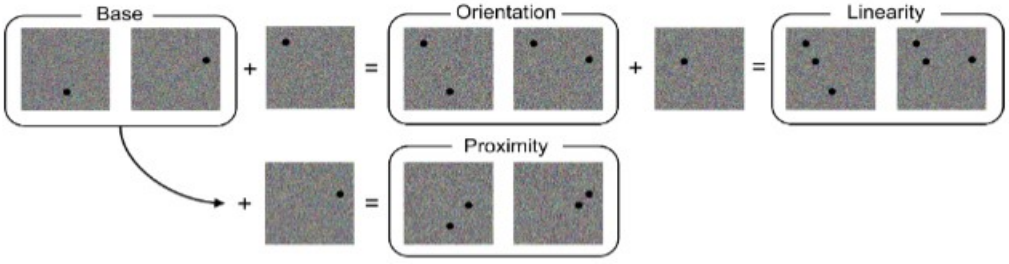


Figure 15. Pomerantz and Portillo (2011) measured Gestalts by constructing a base pair of images (two dots in different locations) and then adding the same context stimulus to each base such that the new image pairs could be distinguished not only using the location of the dots in the base, but also the orientation, linearity, or proximity of the dots. They reported that human participants are faster to distinguish the pair of stimuli in the latter conditions than in the base condition. By contrast, the various DNNs, including DNNs that perform well on Brain-Score, treat the pairs in the orientation, linearity, and proximity conditions as more similar. Images taken from Biscione and Bowers (2022).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

4.2.4 Illusions: Another obvious and striking feature of human vision is the range of visual illusions we experience. There are a few reports that some predictive coding models (e.g., PredNet) display some human-like illusions (e.g., Lotter, Kreiman, & Cox, 2020), although again, more work is needed to determine the extent of the similarity, and most illusions have been given no consideration. By contrast, illusions have been central to the development of theories and models of vision in psychology (most notably by Grossberg; for an excellent and accessible review see Grossberg, 2021) because they provide insight into the way that lightness, color, shape, occlusion, and other stimulus features are used and combined by the human visual system. Interesting, although PredNet captures a number of psychological findings better than standard DNNs, it performs poorly on Brain-Score, currently ranked 177 out of 216 models listed, and Grossberg's models are not even image computable.

4.2.5 Limits in visual short-term memory capacity and attention: There is a variety of evidence suggesting that the visual system attends and encodes approximately four items at a time in short-term memory (Cowan, 2001; Pylyshyn & Storm, 1988; Sperling, 1960). For example, in "multi-object tracking" experiments, multiple dots or objects move around in a display and participants need to track the movement of a subset of them. Participants generally track about four items (Pylyshyn & Storm, 1988). Similarly, limits in visual attention are highlighted in visual search experiments in which response times to targets amongst distractors varies with the visual properties of the target and distractor items (Duncan & Humphreys, 1989; Wolfe et al., 1989; Wolfe, 1994). For example, a search for a target that differs from the distractors by one easily discriminable feature tends to proceed in a parallel fashion with no difference in response time as a function of set size, whereas a search for a target that can only be distinguished from distractors by a conjunction of multiple features tends to take longer as a function of the number of items in the display, suggesting serial attentional processing of the items until the target is found (Triesman & Gelade, 1980). Various manifestations of limited short-term memory and attention can be observed in human object recognition and scene processing, including change blindness where (sometimes large) changes in scenes go unnoticed (Simons & Levin, 1997), and illusory conjunctions in which features of

one object are bound to the features of another (e.g., when briefly flashing an image containing a blue square and red circle, participants will sometimes report seeing a red square and blue circle; Treisman & Schmidt, 1982).

However, there is no analogous visual short-term memory constraint in feedforward DNNs, and we are not aware of any reports that recurrent DNNs manifest any of the human errors that reflect biological visual short-term memory and attention constraints. While some recurrent attention networks (RANs) have attempted to address the problem of serial attentional selection via glimpse mechanisms (Mnih et al., 2014; Ba et al., 2014; Xu et al., 2015), such mechanisms do not provide an account of the influence of item features on processing, nor the associated response time effects.

4.2.6 Selective neuropsychological disorders in vision: The key insight from cognitive neuropsychology is that brain damage can lead to highly selective visual disorders. Perhaps the most well-known set of findings is that acquired dyslexia selectively impairs visual word identification whereas prosopagnosia selectively impairs face identification, highlighting how different systems are specialized for recognizing different visual categories (Farah, 2004). Similarly, lesions can selectively impact vision for the sake of identifying objects vs. vision for sake of action in the ventral and dorsal visual systems, respectively (Goodale, & Milner, 1992). Various forms of visual agnosia have provided additional insights into how objects are identified (Farah, 2004), and different forms of acquired alexia have provided insights into the processes involved in visual word identification (Miozzo & Caramazza, 1998). In addition, selective disorders in motion (Vaina, Makris, Kennedy, & Cowey, 1988) and color perception (Cavanagh et al., 1998) have provided further insights into the organization of the visual system. Few studies have considered whether these selective deficits can be captured in DNNs despite the ease of carrying out lesion studies in networks (for some recent investigations see Hannagan et al., 2021; Ratan Murty et al., 2021).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

4.2.7 Computing shape from non-shape information: Shape is the primary feature that humans rely on when classifying objects, but there are notable examples of recognizing objects based on non-shape features. Classic examples include computing shape from shading (Ramachandran, 1988) and structure from motion (Ullman, 1979). These findings provide important information about how various forms of information are involved and interact in computing shape for the sake of object recognition in humans, but this work has been given little consideration when developing DNNs of vision. For some early work with connectionist networks see Lehky and Sejnowski (1988), and for some recent work with DNNs in this general direction see Fleming and Storrs (2019).

4.2.8 Four correspondences reported by Jacob et al. (2020): As discussed above, Jacob et al. (2020) identified several dissimilarities between DNNs and humans. They also reported four behavioral experiments that they took as evidence of important similarities, but in all cases, the results lend little support for their conclusion and more work is required. First, the authors report that both DNNs and humans respect Weber's Law, according to which the just noticeable difference between two stimuli is a constant ratio of the original stimulus. However, the conditions under which Weber's Law was assessed in humans (reaction times in an eye-tracking study) and DNNs (the Cosine similarity between activation values in hidden layers) were very different, and DNNs only manifest this effect for one of the two stimulus dimensions tested (line lengths but not image intensities). Furthermore, DNNs only supported a Weber's Law effect at the highest convolutional layers, whereas in humans, these effects are the product of early vision (e.g., Van Hateren,1993). Second, Jacob et al. found that DNNs, like humans, are sensitive to scene incongruencies, with reduced object recognition when objects are presented in unusual contexts (e.g., an image of an axe in a supermarket). However, CNNs tend to be far more context-dependent than humans, with DNNs failing to identify objects in unusual contexts, such as an elephant in a living room (Rosenfeld, Zemel, Tsotsos, 2018). Third, Jacob et al. reported that DNNs show something analogous to the Thatcher effect in which humans are relatively insensitive to a specific distortion of a face (the inversion of the mouth) when the entire face is inverted. However, they did not test a key feature of the

Thatcher effect, namely, that it is stronger for faces compared to similar distortions for other categories of objects (Wong, Twedt, Sheinberg, & Gauthier, 2010). Fourth, the authors reported that both humans and DNNs find reflections along the vertical axis (mirror reversals) more similar than reflections along the horizontal axis (inverting an image). However, it is unclear how much weight should be given to this success given that both humans and DNNs experience reflections along the vertical axis much more often. It seems likely that any model that learns could account for this finding.

In sum, many key psychological phenomena have largely been ignored by the DNN community, and the few reports of interesting similarities are problematic or require additional research to determine whether the outcomes reflect theoretically meaningful correspondences or are instead mediated by qualitatively different processes. Furthermore, the few promising results are embedded in a long series of studies that provide striking discrepancies between DNNs and human vision (as summarized above).

5 Deep problems extend to neighboring fields

Although we have focused on DNN models of human vision, the underlying problem is more general. For example, consider DNNs of audition and natural language processing. As is the case with vision, there is excitement that DNNs enable some predictive accuracy with respect to human brain activity (e.g., Kell et al., 2018; Millet et al., 2022; Schrimpf et al., 2021) but at the same time, when models are tested against psychological findings, they fail to support key human-like performance patterns (e.g., Feather et al., 2019; Weerts et al., 2021; Adolfi et al., 2022). And again, the prediction-based experiments used to highlight DNN-human similarities rely on datasets that are not manipulated to test hypotheses about how the predictions are made. For instance, Caucheteux, Gramfort, and King (2022) report that the DNN GPT-2 that generates impressively coherent text also predicts brain activation of humans who listen to 70 min of short stories, with the correlation between the true fMRI responses and the fMRI responses linearly predicted from the model approaching .02 (or approximately .004 of the BOLD variance). In addition,

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

Caucheteux et al. highlight that these predictions correlate with subjects' comprehension scores as assessed for each story at a much higher level ($r = 0.50$, $p < 10{-}15$), and based on this, the authors concluded: "Overall, this study shows how deep language models help clarify the brain computations underlying language comprehension". However, given that the stories were not systematically manipulated to test any hypothesis, this correlation could have other causes, such as the frequency of words in the stories. Indeed, when the correlation between actual BOLD and predicted BOLD is approximately .02, there are undoubtedly many confounding factors that could drive the latter correlation.

Similarly, DNNs that generate coherent text also successfully reproduce a range of human language behaviors, such as accurately predicting number agreement between nouns and verbs (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018). Again, this has led researchers to suggest that DNNs may be models of human linguistic behavior (e.g., Pater, 2019). However, Mitchell and Bowers (2020) show that such networks will also happily learn number agreement in impossible languages within unnatural sentence structures, i.e. structures that are not found within any natural languages and which humans struggle to process. This ability to learn impossible languages is similar to the ability of DNNs to recognize ~1 million instances of random TV-static (Section 4.1.8). In addition, when Mitchell and Bowers (2020) analyzed how knowledge was stored in these networks they found overlapping weights supporting the natural and unnatural structures, again highlighting the non-human-like nature of the knowledge learned by the networks. So again, running controlled experiments that manipulate independent variables highlight important differences between DNNs and humans. It is also important to note that state-of-the-art DNNs of natural language processing receive training that far exceeds any human experience with languages (for example, GPT-2 was trained on text taken from 45 million website links and GPT-3 was trained on 100s of billions of words). This highlights how these DNNs are missing key human inductive biases that facilitate the learning of natural languages but impair the learning of unstructured languages (something akin to a human language acquisition device).

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

Likewise, in the domain of memory and navigation, there are multiple papers claiming that grid cells in the entorhinal-hippocampal circuit emerge in DNNs trained on path integration, that is, estimating one's spatial position in an environment by integrating velocity estimates. This is potentially an important finding given that grid cells in the entorhinal-hippocampal circuit are critical brain structures for navigation, learning and memory. However, it turns out that these results are largely driven by a range of post-hoc implementation choices rather than principles of neural circuits or the loss function(s) they might optimize (Schaeffer et al., 2022). That is, when Schaeffer et al. systematically manipulated the encoding of the target or various hyperparameters, they found the results that were idiosyncratic to specific conditions, and these conditions may be unrealistic. The problem in all cases is that DNN-human similarities are quick to be highlighted and the conclusions are not supported when more systematic investigations are carried out.

6 How should we model human vision?

The appeal of DNNs is that they are an extraordinary engineering success story, with models of object recognition matching or exceeding human performance on some benchmark tests. However, as we have argued, the claim that these models recognize objects in a similar way to humans is unjustified. How can DNNs be useful to scientists interested in modeling human object recognition and vision more broadly? In our view, the first step is to start building models of human object recognition and vision that account for key experimental results reported in psychology rather than ones that perform best on prediction-based experiments. The approach should be the same as it is for all scientific endeavors: use models to test specific hypotheses about how a system works.

6.1 Four different approaches to developing biologically plausible models of human vision:

If one accepts our argument that DNN models of human vision should focus on accounting for experimental studies that manipulate independent variables, it is still the case that very different

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

approaches might be pursued. In our view, all the following approaches should be considered. The simplest transition would be to continue to work with standard DNNs that perform well in identifying naturalistic images but modify their architectures, optimization rules, and training environments to better account for key experimental results in psychology (many of which are reviewed in Section 4) as well as other datasets that assess key behavior findings under controlled conditions (e.g., Crosby, Beyret, & Halina, 2019). This would just involve moving from prediction-based experiments to controlled ones. Key experiments from psychology (as reviewed in Section 4) could be tNote, that the authors of Brain-Score (Schrimpf et al., 2020ab) have highlighted that more benchmarks will be added to the battery of tests, but the problem remains that these and many other authors are making strong claims based on current results, and when experiments are added to the Brain-Score benchmark that do manipulate independent variables (e.g., Geirhos et al., 2021), these manipulations are ignored and the data are analyzed in a prediction-based analysis. Given that current DNNs designed to classify naturalistic images account for almost no psychological findings, it is not clear whether modifications of existing models will be successful, but it is worth exploring, if only to highlight how very different approaches are needed.

Another approach would be to abandon the DNNs that have been built to support engineering objectives (such as performing well on large datasets like ImageNet) and focus on networks designed to account for key psychological phenomena directly. For example, consider the work of Stephen Grossberg and colleagues, recently reviewed in an accessible book that avoids mathematics and focuses on intuitions (Grossberg, 2021). Their models include inhibitory mechanisms designed to support Weber Law dynamics so that networks are sensitive to both small visual contrasts as well as encoding a wide range of visual intensities (the noise-saturation dilemma; Carpenter & Grossberg, 1981); circuits to account for various grouping phenomena that lead to illusory boundaries amongst other illusions (Grossberg & Mingolla,1987); complementary circuits for computing boundaries and surfaces in order to explain the perception of occluded objects, figure-ground organization, and a range of additional visual illusions (Grossberg, 2000); Adaptive Resonance Theory (ART) networks that learn to classify new visual

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

categories quickly without catastrophically forgetting previously learned ones (the stability-plasticity dilemma; Grossberg, 1980); amongst other neural designs used to address core empirical findings. Although these models cannot classify photographic images, they provide more insights into how the human visual system works compared to the DNNs that sit at the top of the Brain-Score leaderboard.

Yet another approach (that overlaps in various ways with the approaches above) would be to build models that support various human capacities that current DNNs struggle with, such as out-of-domain generalization and visual reasoning. That is, rather than making DNNs more human-like in domains in which they are already engineering successes (e.g., modifying DNNs that perform well on ImageNet so that they classify images based on shape rather than texture), instead focus on addressing current performance (engineering) failures (e.g., Francis et al., 2017; George, 2017). For example, one longstanding claim is that symbolic machinery needs to be added to neural networks to support the forms of generalization that humans are capable of (Fodor & Pylyshyn, 1988; Greff, van Steenkiste, & Schmidhuber, 2020; Pinker and Prince, 1988; Marcus, 1998; Holyoak and Hummel, 2000). Interestingly, researchers who have long rejected symbolic models have recently been developing models more in line with a symbolic approach in an attempt to support more challenging forms of visual reasoning and generalization (Sabour, Frosst, & Hinton, 2017; Webb, Sinha, & Cohen, 2021; for some discussion see Bowers, 2017). Indeed, a range of different network architectures have recently been advanced to support more challenging forms of generalization (Doumas, Puebla, Martin, & Hummel, in press; Graves et al., 2016; Mitchell & Bowers, 2021; Vankov & Bowers, 2020) because any model of human vision will ultimately have to support these skills. Of course, it is also necessary to assess whether any successful models perform tasks in a human-like way by testing how well the models explain the results from relevant psychological experiments.

Yet another possible way forward is to use evolutionary algorithms to build neural networks and see if human-like solutions emerge. A key advantage of this approach is that neural network architectures might

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

be evolved that are hard to invent, and indeed, it is sometimes argued that evolutionary algorithms may be the fastest route to building artificial intelligence that rivals human intelligence (e.g., Wang et al., 2020). However, with regards to building models of the human visual system, this approach faces a similar challenge to current DNN modeling, namely, there is no reason to expect the evolved solutions will be similar to human solutions. Indeed, as discussed above, the human visual system is the product of many different and unknown selection pressures applied over the course of millions of years (modification with descent) and it will never be possible to recapitulate all these pressures. So however successful models become within this framework, it cannot be assumed that the evolved solutions will be human-like. Again, the only way to find out will be to test these models on relevant psychological datasets.

Whatever approach one adopts to modelling human object recognition and vision more broadly, the rich database of vision experiments in psychology should play a central role in model development and assessment (for related arguments in the domain of object recognition and classical conditioning see Peters, & Kriegeskorte, 2021, and Bhattasali, Tomov, & Gershman, 2021, respectively; but see Lonnqvist, Bornet, Doerig, & Herzog, 2021 for a different perspective). The approach of comparing models on prediction-based experiments makes sense in the context of building models that solve engineering solutions, but when trying to understand natural systems, the standard methods of science should be adopted: use models to test hypotheses that are evaluated in experiments which manipulate independent variables. By this criterion, models developed in psychology provide superior accounts of human vision than current DNNs that have gathered so much attention.

7 Conclusions

https://doi.org/10.1017/S0140525X22002813 Published online by Cambridge University Press

DNNs outperform all other models on prediction-based experiments carried out on behavioral and brain datasets of object recognition but fail to account for almost all psychological studies of vision. This leads to some obvious questions: Do current prediction-based experiments provide a flawed measure of DNN-human similarity? What have we learned about human visual recognition from DNNs? In what way are DNNs the "best models of human visual object recognition"? In our view, the most obvious explanation for the contrasting results obtained with prediction-based and controlled studies is that prediction-based studies provide a flawed measure of DNN-human correspondences, and consequently, it is unclear what we can learn about human vision by relying upon them, let alone claim DNNs are the best models of biological object recognition.

We suggest that theorists should adopt a more standard research agenda, namely, assess how well models account for a range of data taken from controlled experiments that manipulate independent variables designed to test specific hypotheses. In this context, models are used to explain key empirical findings, and confidence in models grows to the extent that they survive stringent tests designed to falsify them. We have focused on DNN models of object recognition as this is the domain in which the strongest claims have been made but the same considerations apply to all domains of adaptive behavior. In our view, the current prediction-based studies carried out on behavioral and brain datasets are very likely leading us up blind alleys and distracting us from more promising approaches to studying human vision and intelligence more broadly.

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

Conflicts of Interest: None

References:

Adolfi, F., Bowers, J. S., & Poeppel, D. (2022). Successes and critical failures of neural networks in capturing human-like speech recognition. *arXiv preprint arXiv:2204.03740*.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W. S., & Nguyen, A. (2019). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4845-4854).

Alexander, D. M., & Van Leeuwen, C. (2010). Mapping of contextual modulation in the population response of primary visual cortex. *Cognitive Neurodynamics, 4(1)*, 1-24.

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Baker, N., Kellman, P.J., Erlikhman, G., & Lu, H. (2018). Deep convolutional networks do not perceive illusory contours. In *Proceedings of the 40th Annual conference of the cognitive science society, cognitive science society,* Austin, TX (pp. 1310–1315).

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.

Barrett, D., Hill, F., Santoro, A., Morcos, A., & Lillicrap, T. (2018, July). Measuring abstract reasoning in neural networks. In *International conference on machine learning* (pp. 511-520).

Bhattasali, N. X., Tomov, M., & Gershman, S. (2021, June). CCNLab: A Benchmarking Framework for Computational Cognitive Neuroscience. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (Round 1).

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115-147.

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

Biscione, V., & Bowers, J. S. (2021). Convolutional Neural Networks Are Not Invariant to Translation, but They Can Learn to Be. J*ournal of Machine Learning Research, 22,* 1-28.

Biscione, V., & Bowers, J. S. (2022). Learning Online Visual Invariances for Novel Objects via Supervised and Self-Supervised Training. *Neural Networks, 150*, 222-236. https://doi.org/10.1016/j.neunet.2022.02.017.

Biscione, V., & Bowers, J. S. (2022). Do DNNs trained on natural images acquire gestalt properties?. arXiv preprint arXiv:2203.07302.

Blything, R., Biscione, V., & Bowers, J. (2020). A case for robust translation tolerance in humans and CNNs. A commentary on Han et al. *arXiv* preprint arXiv:2012.05950.

Blything, R., Biscione, V., Vankov, I. I., Ludwig, C. J. H., & Bowers, J. S. (2021). The human visual system and CNNs can both support robust online translation tolerance following extreme displacements. *Journal of Vision, 21(2):9*, 1–16. https://doi.org/10.1167/jov.21.2.9.

Bowers, J.S. (2017). Parallel Distributed Processing Theory in the Age of Deep Networks. *Trends in Cognitive Science, 21*, 950-961.

Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138, 389-414*. DOI: 10.1037/a0026450

Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin, 138,* 423– 426. DOI: 10.1037/a0027750

Bowers, J.S. & Jones, K.W. (**2007**). Detecting objects is easier than categorizing them. *Quarterly Journal of Experimental Psychology, 61*, 552-557.

Bowers, J. S., Vankov, I. I., & Ludwig, C. J. (2016). The visual system supports online translation invariance for object identification. *Psychonomic Bulletin & Review, 23*, 432-438.

Burgess, C. P., et al. (2019). Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:1901.11390.

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.

Cao, Y., Grossberg, S., & Markowitz, J. (2011). How does the brain rapidly learn and reorganize view-invariant and position-invariant object representations in the inferotemporal cortex? *Neural Networks*, *24*(10), 1050-1061.

Carpenter, G. A., & Grossberg, S. (1981). Adaptation and transmitter gating in vertebrate photoreceptors. *Journal of Theoretical Neurobiology*, *1*(1), 1-42.

Caucheteux, C., Gramfort, A., & King, J. R. (2022). Deep language algorithms predict semantic comprehension from brain activity. Scientific Reports, 12(1), 1-10.

Cavanagh, P., Hénaff, M. A., Michel, F., Landis, T., Troscianko, T., & Intriligator, J. (1998). Complete sparing of high-contrast color input to motion perception in cortical color blindness. *Nature Neuroscience, 1(3),* 242-247.

Clune, J., Mouret, J. B., & Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society b: Biological Sciences, 280(1755),* 20122863.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological measurement*, *20*(1), 37-46.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24(1),* 87-114.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A.(2016). Comparison of deep neural networks to spatiotemporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific. Reports, 6*, 1-13.

Crosby, M., Beyret, B., & Halina, M. (2019). The animal-ai olympics. Nature Machine Intelligence, 1(5), 257-257.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Advances in Neural Information Processing Systems, 33*, 13073-13087.

Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000346

Driver, J., & Baylis, G. C. (1996). Edge-assignment and figure–ground segmentation in short-term visual matching. *Cognitive Psychology, 31(3),* 248-306.

Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C. P., Lerchner, A., & Higgins, I. (2019). Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:*1905.12614.

Dujmović, M., Bowers, J.S., Adolfi, F., & Malhotra, G. (2022). The pitfalls of measuring representational similarity using representational similarity analysis. *arXiv preprint* https://www.biorxiv.org/content/10.1101/2022.04.05.487135v1

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

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality. *bioRxiv.*

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychological Review, 124(6),* 740-761.

Evans, B. D., Malhotra, G., & Bowers, J. S. (2022). Biological convolutions improve DNN robustness to noise and generalisation. *Neural Networks, 148*, 96- 110. https://doi.org/10.1016/j.neunet.2021.12.005

Farah, M. J. (2004). *Visual agnosia.* MIT press.

Feather, J., Durango, A., Gonzalez, R., & McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 10078–10089).

Fleming, R. W., & Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences*, *30*, 100-108.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28(1-2),* 3-71.

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review, 124(4),* 483-504.

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S., & Bethge, M. (2021). Five

points to check when comparing visual perception in humans and machines. *Journal of Vision, 21(3):*16, 1-23.

Garrigan, P., & Kellman, P. J. (2008). Perceptual learning depends on perceptual constancy. *Proceedings of the National Academy of Sciences*, *105*(6), 2248-2253.

Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science, 2*, 377–396.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2(11),* 665-673.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *arXiv preprint* arXiv:2006.16736.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *arXiv preprint* arXiv:2106.07411.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, https://arxiv.org/abs/1811.12231

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 7538-7550).

George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., ... & Phoenix, D. S. (2017). A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science, 358*(6368).

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15(1),* 20-25.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature, 538(7626),* 471-476.

Greff, K., et al.. (2019, May). Multi-object representation learning with iterative variational inference. In International Conference on Machine Learning (pp. 2424-2433).

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv preprint* arXiv:2012.05208.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138(3)*, 415–422. https://doi.org/10.1037/a0026884

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16(2),* 152-160.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1-51.

Grossberg, S. (2000). The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences*, *4*, 233-246.

Grossberg, S. (2021). *Conscious Mind, Resonant Brain: How Each Brain Makes a Mind*. Oxford University Press.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological Review, 92(2),* 173-211.

Grossberg, S., & Mingolla, E. (1987). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. In *The adaptive brain II* (pp. 143-210). Elsevier.

Guest, O., & Martin, A. E. (2021). On logical inference over brains, behaviour, and artificial neural networks. *arXiv preprint* psyarxiv.com/tbmcg, doi:10.31234/osf.io/tbmcg

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018, June). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL 2018* (pp. 1195–1205). New Orleans, Louisiana: ACL.

Hacker, C., & Biederman, I. (2018). The invariance of recognition to the stretching of faces is not explained by familiarity or warping to an average face. *arXiv preprint* https://doi.org/10.31234/osf.io/e5hgx

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).

Heinke, D., Wachman, P., van Zoest, W., & Leek, E. C. (2021). A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision. *Vision Research*, *189*, 81-92.

Hermann, K. L., Chen, T., & Kornblith, S. (2019). The origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv*:1911.09071.

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, *75*(4), 624-628.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines.* (pp. 229–264). Cambridge, MA: MIT Press.

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2022). The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. arXiv preprint arXiv:2205.10144.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Deitrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Mahwah, NJ: Erlbaum

Hummel, J.E. (2013). Object recognition. In D. Reisburg (ed.), *Oxford Handbook of Cognitive Psychology*, pp. 32-46. Oxford: Oxford University Press.

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

Izhikevich E. M. (2004) Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks* Volume 15 Issue 5 September 2004 pp 1063–1070 https://doi.org/10.1109/TNN.2004.832719

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1-14.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron, 98*(3), 630-644.e16.

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

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports, 6(1)*, 1-24.

Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex. *Journal of Neurophysiology, 97*, 4296–4309. doi:10.1152/jn.00024.2007

Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 1-13.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. *arXiv preprint arXiv:1906.01558*.

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., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60(6)*, 1126-1141.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4, 1-28.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology, 12(4)*, e1004896.

Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., Rajalingham, R., ... & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *arXiv preprint* arXiv:1909.06161.

Lehky, S. R., & Sejnowski, T. J. (1988). Network model of shape-from-shading: neural function arises from both receptive and projective fields. *Nature, 333*(6172), 452-454.

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, *19*(2), 189-223.

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence, 2(4)*, 210-219.

Lissauer, SH. (1890). Ein Fall von Seelenblindheit nebst einem Beitrage zur Theorie derselben. Archiv für Psychiatrie und Nervenkrankheiten, 21(2), 222-270.

Lonnqvist, B., Bornet, A., Doerig, A., & Herzog, M. H. (2021). A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities. *Journal of Vision, 21*(10), 17-17. doi:https://doi.org/10.1167/jov.21.10.17

Mack, M. L., Gauthier, I., Sadr, J., & Palmeri, T. J. (2008). Object detection and basic-level categorization: Sometimes you know it is there before you know what it is. *Psychonomic Bulletin & Review*, *15*(1), 28-35.

Macpherson, T., Churchland, A., Sejnowski, T., DiCarlo, J., Kamitani, Y., Takahashi, H., & Hikida, T. (2021). Natural and Artificial Intelligence: A brief introduction to the interplay between AI and neuroscience research. *Neural Networks, 144*, 603-613.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience, 35(39),* 13402-13418.

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research, 174*, 57-68.

Malhotra, G., Dujmovic, M., & Bowers, J. S. (2022). Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Computational Biology, 18, e1009572.* https://doi.org/10.1101/2021.10.20.465074

Malhotra, G., Dujmovic, M., Hummel, J., & Bowers, J. S. (2021). The contrasting shape representations that support object recognition in humans and CNNs. *arXiv preprint* https://doi.org/10.1101/2021.12.14.472546

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37(3),* 243-282.

Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.

Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv* preprint arXiv:2002.06177.

Marques, T., Schrimpf, M., & DiCarlo, J. J. (2021). Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. *arXiv preprint* https://doi.org/10.1101/2021.03.01.433495

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*, Henry Holt and Co. Inc., New York, NY, 2(4.2).

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel Distributed Processing* (Vol. 2). Cambridge, MA: MIT press.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 1-12.

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).

Messina, N., Amato, G., Carrara, F., Gennaro, C., & Falchi, F. (2021). Solving the same-different task with convolutional neural networks. Pattern Recognition Letters, 143, 75-80.

Millet, J., & King, J.-R. (2021). Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *ArXiv:2103.01032 [Cs, Eess, q-Bio]*.

Miozzo, M., & Caramazza, A. (1998). Varieties of pure alexia: The case of failure to access graphemic representations. *Cognitive Neuropsychology, 15(1-2)*, 203-238.

Mitchell, J., & Bowers, J. (2020, December). Priorless Recurrent Networks Learn Curiously. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5147-5158).

Mitchell, J., & Bowers, J. S. (2021). Generalisation in Neural Networks Does not Require Feature Overlap. *arXiv preprint arXiv*:2107.06872.

Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* (pp. 2204-2212).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518(7540),* 529-533.

Montero, M.L., Bowers, J.S., Ludwig, C. J., Costa, R. P., Malhotra, G. (2022). Lost in Latent Space: Disentangled Models and the Challenge of Combinatorial Generalisation. *arXiv preprint:* http://arxiv.org/abs/2204.02283

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., & Bowers, J. (2020, September). The role of disentanglement in generalisation. In *International Conference on Learning Representations.*

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception*, *18*, 55-68.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 427-436).

Palmer, S. E. (1999). Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences, 22(6)*, 923-943.

Palmer, S. E. (2003). Visual perception of objects. *Handbook of Psychology*, 177-211.

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *arXiv preprint* arXiv:2102.01955.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language, 95(1):*e41–e74.

Pepperberg, I. M., & Nakayama, K. (2016). Robust representation of shape in a Grey parrot (Psittacus erithacus). *Cognition, 153*, 146-160.

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 1-18.

Pessoa, L., Thompson, E., & Noë, A. (1998). Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences, 21(6)*, 723-748

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42(8)*, 2648-2669.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193.

Pomerantz J.R., & Portillo, M.C. (2011). Grouping and Emergent Features in Vision: Toward a Theory of Basic Gestalts. *Journal of Experimental Psychology: Human Perception and Performance, 10*, *37*, 1331–1349. 10.1037/A0024330

Pomerantz, J. R., Sager, L. C., & Stoever, R. J. (1977). Perception of wholes and of their component parts: some configural superiority effects. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 422- 435. https://doi.org/10.1037/0096-1523.3.3.422

Popper, K. (2005). The logic of scientific discovery. Routledge.

Puebla, G., & Bowers, J. S. (2022). Can deep convolutional neural networks support relational reasoning in the same-different task?. *Journal of Vision, 22*(10), 11-11.

Pylyshyn, Z. W. & Storm, R. W. (1988) Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision 3,* 179–97.

Raizada, R., & Grossberg, S. (2001). Context-sensitive bindings by the laminar circuits of V1 and V2: A unified model of perceptual grouping, attention, and orientation contrast. *Visual Cognition, 8*, 431– 466.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 38*(33), 7255-7269.

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience, 35(35),* 12127-12136.

Ramachandran, V. S. (1992). Filling in gaps in perception: Part I. *Current Directions in Psychological Science*, *1*(6), 199-205.

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature*, *331*(6152), 163-166.

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications, 12(1),* 1-14.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... & Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience, 22(11)*, 1761-1770.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017, July). Cognitive psychology for deep neural networks: A shape bias case study. In International conference on machine learning (pp. 2940-2949).

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological review, 107(2), 358-367/ doi: 10.1037/0033-295x.107.2.358.

Rosenfeld A., Zemel R., Tsotsos J.K. (2018). The elephant in the room. *arXiv preprint* arXiv:1808.03305

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, *9*(2):5, 1-11.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv preprint* arXiv:1710.09829.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems, 30*, 4967–4976.

Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. bioRxiv.

Schott, L., von Kügelgen, J., Träuble, F., Gehler, P., Russell, C., Bethge, M., ... & Brendel, W. (2021). Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv*:2107.08221.

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).

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... & DiCarlo, J. J. (2020a). Brain-score: Which artificial neural network for object recognition is most brain-like? *arXiv preprint* https://doi.org/10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020b). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron, 11*, 413-423.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., & Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *arXiv preprint* arXiv:2006.07710.

Silver, D. et al. A general reinforcement learning algorithm that masters chess, shogi and Go through self-play. *Science 362*, 1140–1144 (2018).

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261-267.

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

Sperling, G. (1960) The information available in brief visual presentations. *Psychological Monographs 74:* Whole No. 498.

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

Stanley, K. O., Clune, J., Lehman, J., & Miikkulainen, R. (2019). Designing neural networks through neuroevolution. *Nature Machine Intelligence, 1(1),* 24-35.

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, 33(10)*, 2044-2064.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107-141.

Tsvetkov, C., Malhotra, G., Evans, B., & Bowers, J. (2020). Adding biological constraints to deep neural networks reduces their capacity to learn unstructured data. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* 2020, Toronto, Canada.

Tsvetkov, C., Malhotra, G., Evans, B., & Bowers, J (in press). The Role of Capacity Constraints in Convolutional Neural Networks for Learning Random Versus Natural Data. *Neural Networks.* https://doi.org/10.1101/2022.03.31.486580

Truzzi, A., & Cusack, R. (2020). Convolutional neural networks as a model of visual activity in the brain: Greater contribution of architecture than learned weights. Bridging AI and Cognitive Science. *International Conference on Learning Representations.*

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are Convolutional Neural Networks or Transformers more like human vision?. arXiv preprint arXiv:2105.07197.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *203*(1153), 405-426.

Ullman, S., and Basri, R. (1991). Recognition by Linear Combination of Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13*, No. 10, pp. 992-1006.

Vaina, L. M. Makris, N., Kennedy, D., & Cowey, A. (1988). The selective impairment of the perception of first-order motion by unilateral cortical brain damage. *Visual Neuroscience, 15,* 333–348.

Vankov, I. I., & Bowers, J. S. (2020). Training neural networks to encode symbols enables combinatorial generalization. *Philosophical Transactions of the Royal Society B, 375*(1791), 20190309.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172-1217.

Wang, R., Lehman, J., Rawal, A., Zhi, J., Li, Y., Clune, J., & Stanley, K. (2020, November). Enhanced POET: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *International Conference on Machine Learning* (pp. 9940-9951).

Wang J, Zhang Z, Xie C, Zhou Y, Premachandran V, Zhu J, Xie L, Yuille A (2018) Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications 2(3):*4

Webb, T. W., Sinha, I., & Cohen, J. D. (2021). In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2012.14601

Wertheimer, Max (1912) Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie* (in German), 61, 161–265.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419-433.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202-238.

Wong, Y. K., Twedt, E., Sheinberg, D., & Gauthier, I. (2010). Does Thompson's Thatcher effect reflect a face-specific mechanism? *Perception, 39(8),* 1125-1141.

Woolley, B. G., & Stanley, K. O. (2011, July). On the deleterious effects of a priori objectives on evolution and representation. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation* (pp. 957-964).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048-2057). PMLR.

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications, 12*(1), 1-16.

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.

Young, T., 1802. Bakerian Lecture: On the Theory of Light and Colours. Philosophical Transactions of the Royal Society London, 92:12–48. doi: 10.1098/rstl.1802.0004

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications, 10(1),* 1-7.

Zhang, R. (2019, May). Making convolutional networks shift-invariant again. In International conference on machine learning (pp. 7324-7334). *Proceedings of Machine Learning Research*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In 5th *International Conference on Learning Representations*, Toulon, France, April 24-26.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1-9.

Zhu, H., Tang, P., Park, J., Park, S., & Yuille, A. (2019). Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*.

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).

Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055