Behavioral and Brain Sciences

cambridge.org/bbs

Deep problems with neural network models of human vision

Jeffrey S. Bowers[a], Gaurav Malhotra[a], Marin Dujmović[a], Milton Llera Montero[a], Christian Tsvetkov[a], Valerio Biscione[a], Guillermo Puebla[a], Federico Adolfi[a,b], John E. Hummel[c], Rachel F. Heaton[c], Benjamin D. Evans[d], Jeffrey Mitchell[d] and Ryan Blything[e]

[a]School of Psychological Science, University of Bristol, Bristol, UK; [b]Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt am Main, Germany; [c]Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL, USA; [d]Department of Informatics, School of Engineering and Informatics, University of Sussex, Brighton, UK and [e]School of Psychology, Aston University, Birmingham, UK
j.bowers@bristol.ac.uk; https://jeffbowers.blogs.bristol.ac.uk/
gaurav.malhotra@bristol.ac.uk
marin.dujmovic@bristol.ac.uk
m.lleramontero@bristol.ac.uk
christian.tsvetkov@bristol.ac.uk
valerio.biscione@gmail.com
guillermo.puebla@bristol.ac.uk
fedeadolfi@gmail.com
jehummel@illinois.edu
rmflood2@illinois.edu
b.d.evans@sussex.ac.uk
j.mitchell@napier.ac.uk
r.blything@aston.ac.uk

Target Article

Cite this article: Bowers JS *et al*. (2023) Deep problems with neural network models of human vision. *Behavioral and Brain Sciences* **46**, e385: 1–77. doi:10.1017/S0140525X22002813

Target Article Accepted: 11 November 2022
Target Article Manuscript Online: 1 December 2022
Commentaries Accepted: 22 March 2023

Keywords:
Brain-Score; computational neuroscience; deep neural networks; human vision; object recognition

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 22) and an Authors' Response (p. 67). See bbsonline.org for more information.

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 modeling approaches that focus on psychological data.

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; Erdogan & Jacobs, 2017; Hummel & Biederman, 1992; Marr, 1982; 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

© The Author(s), 2022. Published by Cambridge University Press

Jeffrey S. Bowers is a professor of Psychology in the School of Psychological Science, University of Bristol. His research exploits behavioral, neuropsychological, and computational methods to study the representations and processes that support human perception, memory, and language. He also is interested in relating basic research into human cognition to pedagogy, with a specific interest in reading instruction.

Gaurav Malhotra is a research fellow at the University of Bristol. He earned his PhD from the University of Edinburgh in 2009. His research combines computational models with behavioral experiments to understand how our environment and biology shape our cognition. His main areas of interest are human vision and decision making.

Marin Dujmović is a PhD student at the School of Psychological Science, University of Bristol. As a cognitive psychologist his main interests are understanding basic processes in the areas of reasoning and rationality as well as human vision. His work has focused on comparing computational models and human cognition by employing both simulations and behavioral experiments.

Milton Llera Montero is a PhD student at the School of Psychological Science and the Computational Neuroscience Unit in the Department of Computer Science, University of Bristol. His main area of interest is in compositional generalization in vision and the mechanisms that support such abilities in both humans and AIs. He is also interested in decision-making models for naturalistic stimuli and their relation to findings in neuroscience and psychology.

Christian Tsvetkov is a PhD student at the School of Psychological Science, University of Bristol. His research interests include the study of information processing and learning in biological and artificial systems, particularly in the domain of vision. His work focuses on modeling biologically inspired mechanisms and computational constraints using neural networks in order to investigate properties of generalization and perceptual organization.

Valerio Biscione is a research associate in the School of Psychological Science at the University of Bristol. His research focuses on similarities and differences between the human visual system and modern artificial networks, especially regarding invariance to object transformations and Gestalt grouping. He is also interested in spiking networks, bio-plausible models, and evolutionary mechanisms for the emergence of complex behavior.

Guillermo Puebla is a Postdoc at the National Center for Artificial Intelligence CENIA. His research focuses on relational learning and reasoning both in humans and computational models, such as artificial neural networks. He studies these topics through simulations and behavioral studies in humans.

Federico Adolfi is a PhD student at the University of Bristol and the Ernst-Strüngmann Institute for Neuroscience, Max-Planck Society. He works on developing and evaluating computational models as explanations for natural and artificial cognition, and uses tools from theoretical computer science to expose their complexity-theoretic properties. He is also interested in how we come up with such explanations, what makes them plausible, and what makes them useful. He received the Computational Modeling award and the Disciplinary Diversity and Integration prize from the Cognitive Science Society for work combining theoretical computer science and cognitive science.

John E. Hummel is a professor of Psychology and Philosophy at the University of Illinois at Urbana–Champaign. He studies shape perception, object recognition, concepts, and reasoning, both empirically and by building empirically and neurally constrained computational models of these processes. He is especially interested in how neural computing architectures give rise to symbolic thought, with a focus on the representation and processing of explicitly relational representations.

Rachel F. Heaton is a doctoral candidate in the Department of Psychology at the University of Illinois Urbana–Champaign. She earned a Masters of Fine Arts in Industrial Design from the University of Illinois Urbana–Champaign. She develops neurally plausible computational models of attention and visual cognition, and is interested in the way that representations and processes in visual reasoning relate to the user experience of designed objects.

Benjamin D. Evans is a lecturer in the Department of Informatics at the University of Sussex. His research spans computational neuroscience, machine learning and bio-inspired AI, particularly in the domain of visual information processing. He employs a variety of modeling paradigms including spiking neural networks and deep-learning models, which he constrains and enhances with the inductive biases of their biological counterparts to improve their robustness and generalization.

Jeffrey Mitchell is a lecturer in Informatics at the University of Sussex. His research covers a number of topics in computational linguistics and cognitive science. He is interested in understanding the generalization abilities of humans and machines.

Ryan Blything is a lecturer in Psychology at Aston University and earned his PhD from the University of Manchester. His research investigates the capacity for generalization in humans and artificial neural networks, as well as the representations needed to support these generalizations, using the domains of visual object recognition and psycholinguistics as test cases.

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% 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, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021; Zhuang et al., 2021).

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 modeling 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.

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

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 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…. (Zhuang 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, Kietzmann, Walther, Mehrer, & Kriegeskorte, 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)

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., 2019)
>
> 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 blood-oxygen-level-dependent [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 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. Due to 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 versus line drawings of the same objects under the same viewing conditions to assess the role of shape versus texture/color in object identification (see sect. 4.2.3).

To illustrate the prediction-based nature of these studies, consider the image dataset from Kiani, Esteky, Mirpour, and Tanaka (2007) used by Khaligh-Razavi and Kriegeskorte (2014) to assess how well DNNs can predict single-cell responses in macaques and fMRI BOLD signals in humans using representational similarity analysis (RSA). This dataset includes objects from six categories (see Fig. 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 to make inferences regarding the similarity of DNNs to the human (or macaque) visual object recognition system.

Or consider the Brain-Score benchmark that includes a range of behavioral and brain studies that together are used to rate a

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

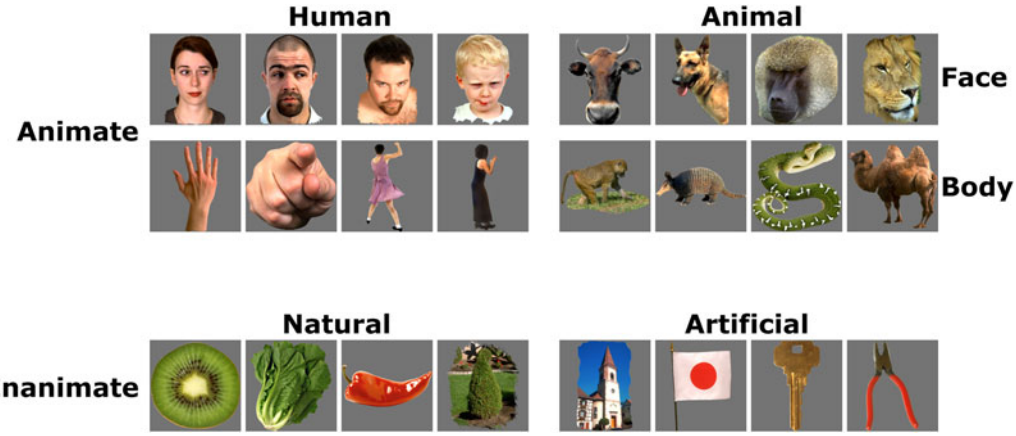


Figure 1. Example images from Kiani et al. (2007) that include images from six categories.

model's similarity to human object recognition (Schrimpf et al., 2020a, 2020b). Currently five studies are used to assess how well DNNs predict brain activation in inferotemporal (IT) cortex. The first of these (Majaj, Hong, Solomon, & DiCarlo, 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.

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 Fig. 2; Rajalingham, Schmidt, & DiCarlo, 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 sect. 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 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, Spoerer, Kriegeskorte, & Kietzmann, 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

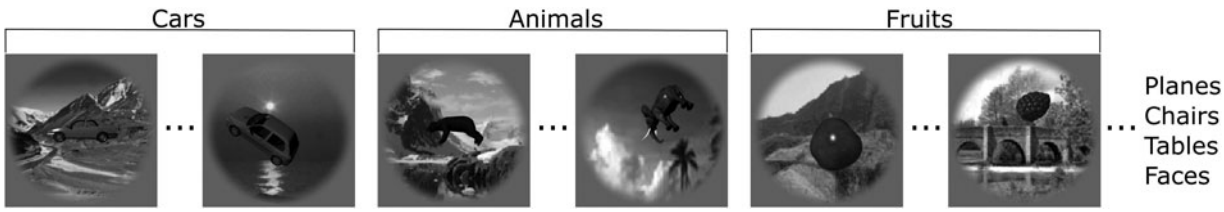


Figure 2. Example images of cars, fruits, and animals at various poses with random backgrounds from Majaj et al. (2015).

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

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 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 (2023) 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 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 "shortcuts" that humans never rely on (e.g., Geirhos et al., 2018; Malhotra, Dujmovic, & Bowers, 2022; Malhotra, Evans, & Bowers, 2020; 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 shortcuts, 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 shortcuts also perform well on prediction-based experiments. Dujmović, Bowers, Adolfi, and Malhotra (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 Fig. 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., Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Kiat et al., 2022; Kriegeskorte, Mur, & Bandettini, 2008a; Kriegeskorte et al., 2008b).

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 et al. (2015) that contained a pixel patch confound (call it DNN-pixel) as well as unperturbed photos (DNN-standard), similar to the Malhotra et al.'s (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 pretrained on naturalistic stimuli (ImageNet dataset) and fine-tuned on the unperturbed images (Fig. 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 RSAs.

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., 2020a, 2020b). 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

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

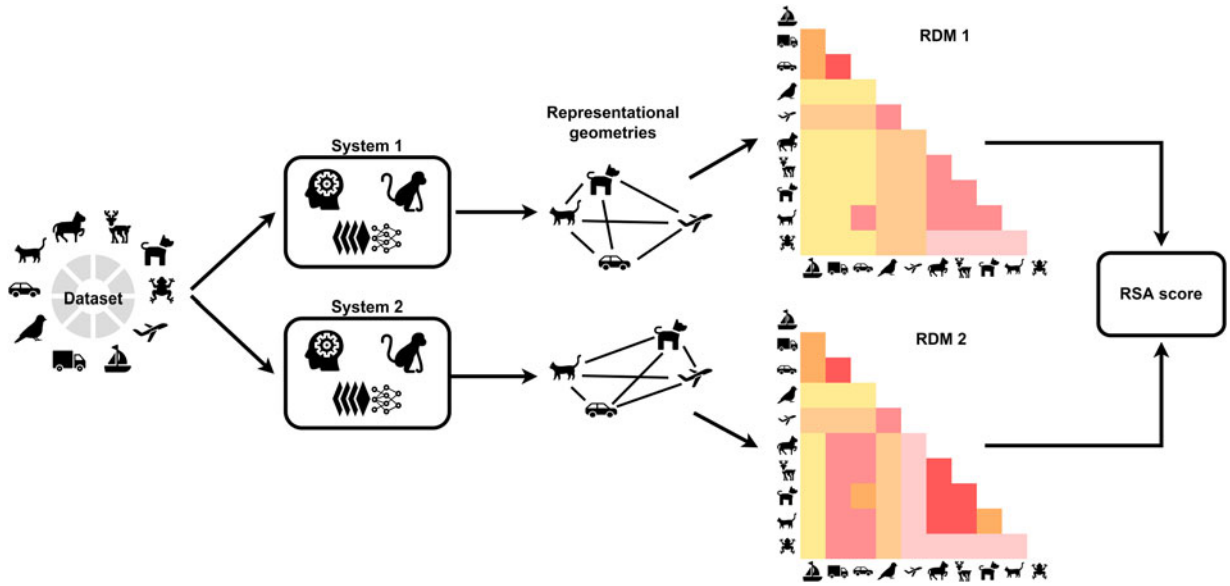


Figure 3. RSA calculation. A series of stimuli from a set of categories (or conditions) are used as inputs to two different systems (e.g., a brain and a DNN). The corresponding neural or unit activity for each stimulus is recorded and pairwise distances in the activations within each system are calculated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

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 sect. 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 this comes from ongoing work by Dujmović, Bowers, Adolfi, and Malhotra (2023) 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 Fig. 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 (shortcuts) 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 and Bonner (2022) have shown that effective latent dimensionality of DNNs significantly

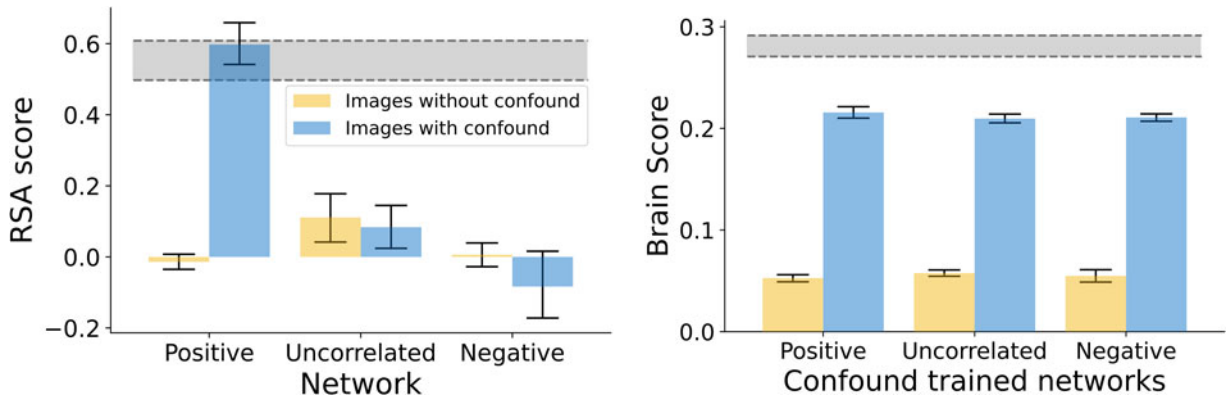


Figure 4. RSA (left) and Brain-Score (right) for networks trained on predictive pixels. The location of the pixel patches varied across conditions, such that the location was positively, negatively, or uncorrelated with the representational distances between classes in the macaque IT. When the pixel distances are positively correlated in the training set, RSA scores approached scores achieved by networks pretrained 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 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).

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

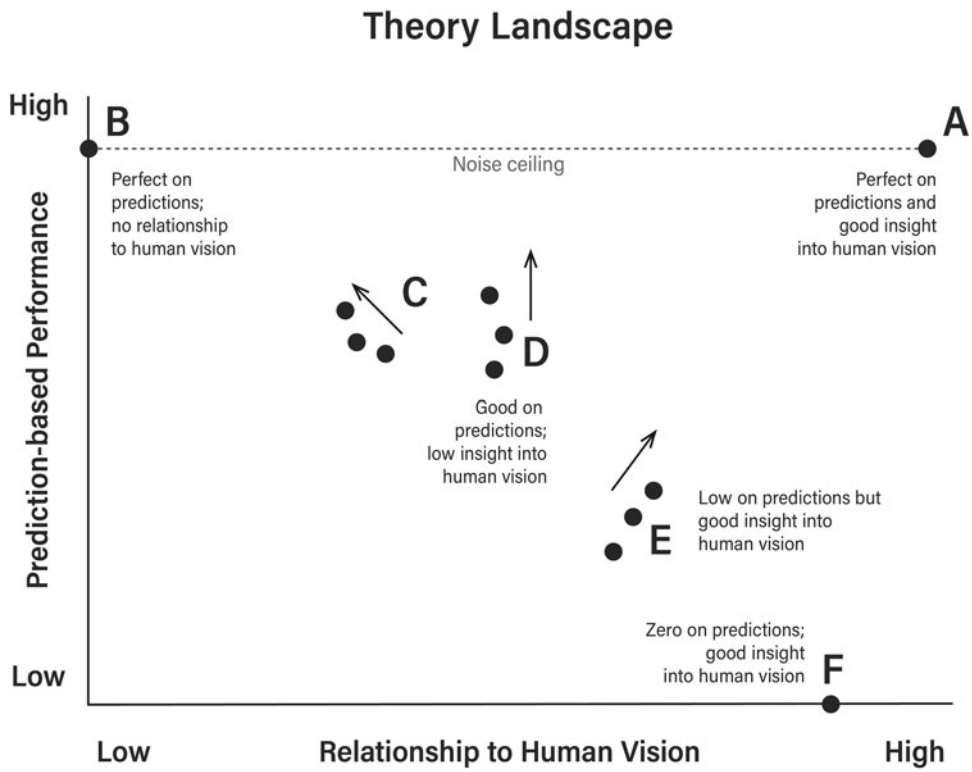


Figure 5. Different models fall in different parts of the theory landscape. Critically, it is possible to do well on prediction-based experiments despite poor correspondences to human vision, and there is no reason to expect that modifying a model to perform better on these experiments will necessarily 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.

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, pretraining, training paradigm, and so on 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 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.

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 cat and a texture of an elephant is classified as a cat; Geirhos et al., 2019; for more details see sect. 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 et al., 2021), or that it is important to add top-down connections that characterize the human visual system (e.g., Zhuang et al.,

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

2021), and so on. 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 photo-realistic 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 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 any different from the current DNNs given that they characterize 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 sect. 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 convolutional neural networks (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 regard 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 noted above, the Brain-Score match to high-level vision in IT is based on five studies that rely on a total of three monkeys presented with two very similar image datasets. Similarly, the reports of high RSAs between DNNs and human vision have 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 only a 0.42 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 0.783 on one V1 dataset (ranking in the top 10) and 0.245 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 et al., 2015, 2018; Tuli, Dasgupta, Grant, & Griffiths, 2021). Geirhos, Meding, and Wichmann (2020b) 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, 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 0.17. Clearly, the different methods used to measure

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

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

Apart from the principled and practical problems with prediction-based experiments, the general approach of modeling 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, Dujmovic, Hummel, and Bowers (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 sect. 4.2.3) and shape-processing biases (see sect. 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 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. Under 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 counterintuitive 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, Clune, Lehman, & Miikkulainen, 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 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, 2012a, 2012b; 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, it is not only 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

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



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).

best models of human vision and object recognition. Rather, they imply that the 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 (i.e., 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 Fig. 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), 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.

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

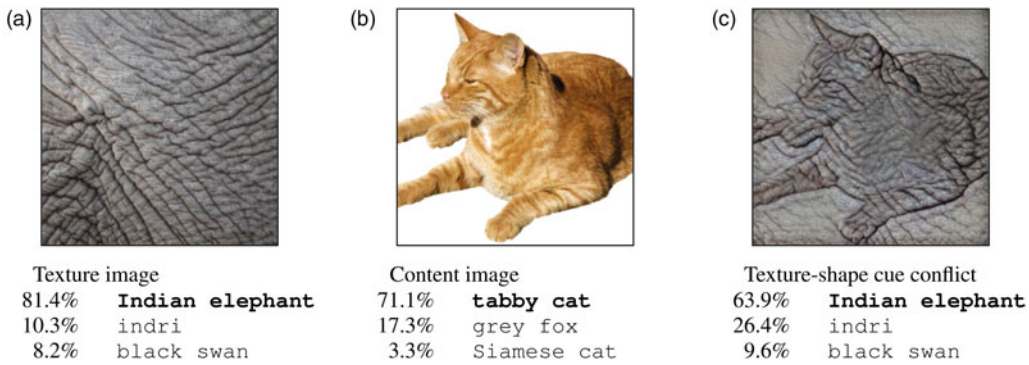


Figure 7. Illustration of a style-transfer image in which (a) the texture of an elephant and (b) the shape of a cat that combine to form (c) the shape of a cat with the texture of an elephant. The top three classifications of a DNN to the three images are listed below each image, with the model classifying 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).

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

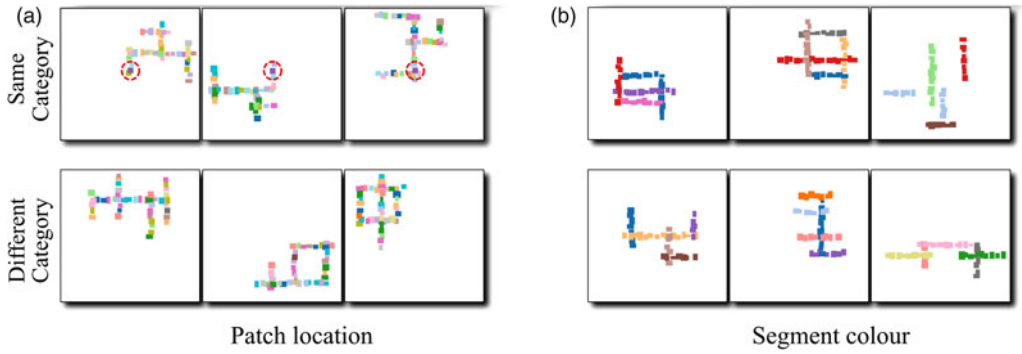


Figure 8. Examples of novel stimuli defined by shape as well as one other nonshape feature. In (a) global shape and location of one of the patches define a category, and for illustration, the predictive patch is circled. Stimuli in the same category (top row) have a patch with the same color and the 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 nonshape feature (red segment) but no shape feature. Images taken from Malhotra et al. (2022).

an elephant) to assess the relative importance of texture versus 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 Fig. 7). Indeed, the CORnet-S model described as one of the best models of human vision largely classifies objects based on texture (Geirhos et al., 2020b), and this contrast between DNNs and humans extends to children and adults (Huber, Geirhos, & Wichmann, 2022; but see Ritter, Barrett, Santoro, & Botvinick, 2017, for the claim that DNNs have a human-like shape-bias).

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 nonshape diagnostic feature (including patch location and segment color as shown in Fig. 8). Humans showed a strong shape-bias when classifying these images, and indeed, could not learn to classify the objects based on some nonshape features. By contrast, DNNs had a strong bias to rely on these very same nonshape features. Importantly, when the DNNs were pretrained 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 nonshape features when subsequently trained to classify these stimuli. This was the case even after freezing the convolutional layers of a shape-biased ResNet50 (i.e., 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, 2020).

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., 2020) or architecture (Evans, Malhotra, & Bowers, 2022) 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 to a large body of research showing that humans tend to rely on the global shape of objects, Baker, Lu, Erlikhman, and Kellman (2018b) 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 functioned 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 Fig. 9).

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 Fig. 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.

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 (Livingstone & 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

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

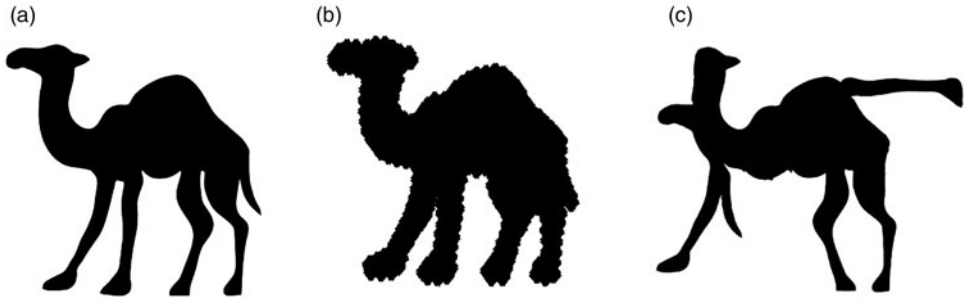


Figure 9. Illustration of (a) a silhouette image of a camel, (b) and image of a camel in which local shape features were removed by including jittered contours, and (c) and image of a camel in which global shape was disrupted. The DNNs had more difficulty under conditions (b) than (c). Images taken from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

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 convolutional neural networks (CNNs) may also help explain why DNNs do not have a strong shape bias when classifying objects.

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 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.

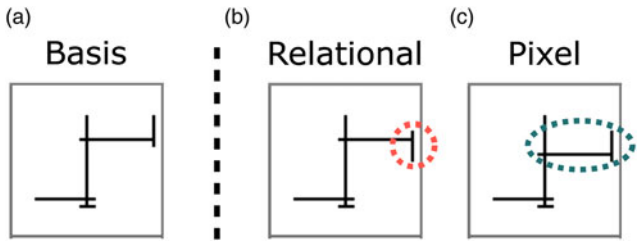


Figure 10. Example of (a) a basis object, (b) a relational variant object that was identical to the basis object except that one line was moved so that its "above/below" relation to the line to which it was connected changed (from above to below or vice-versa), as highlighted by the circle, and (c) a pixel variant object that was identical to the basis object except that 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).

4.1.7. 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.,

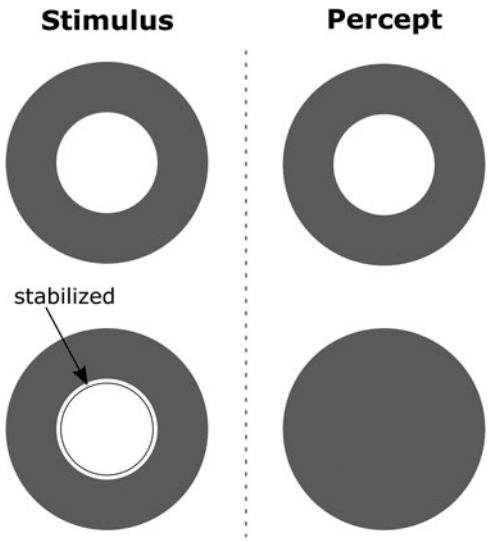


Figure 11. Phenomenon of filling-in suggests that edges and textures are initially processed separately and then combined to produce percepts. In this classic example from Krauskopf (1963), an inner green disk (depicted in white) is surrounded by a red annulus (depicted in dark gray). Under normal viewing 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).

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

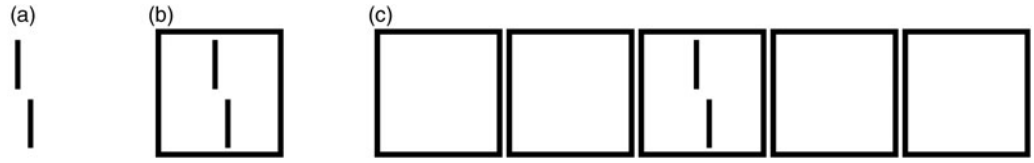


Figure 12. (a) Under the standard vernier discrimination conditions two vertical lines are offset, and the task of the participant is to judge whether the top line is to the left or right of the bottom line. (b) Under the crowding condition the vernier stimulus is surrounded by a square and discriminations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

2018; Zhu, Tang, Park, Park, & Yuille, 2019). It should be noted that the larger DNNs do better on degraded images (e.g., CLIP trained on 400 million 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 superhuman level. For example, Zhang, Bengio, Hardt, Recht, and Vinyals (2017) trained standard DNNs with ~1 million images composed of random pixel activations (TV static-like images) that were randomly assigned to 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, 2023) 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 judgments for novel three-dimensional (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., 2020a). However, similarity judgments break down for unfamiliar objects. For example, German and Jacobs (2020) measured human similarity judgments 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 relations between object parts (Malhotra et al., 2021), a likely factor in the human similarity judgments 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 regard 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; Zhao, Zheng, Xu, & Wu, 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 collinearity, curvature, cotermination, and so on 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 judgments (Funke et al., 2021; Messina, Amato, Carrara, Gennaro, & Falchi, 2021) the models were 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 Fig. 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

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

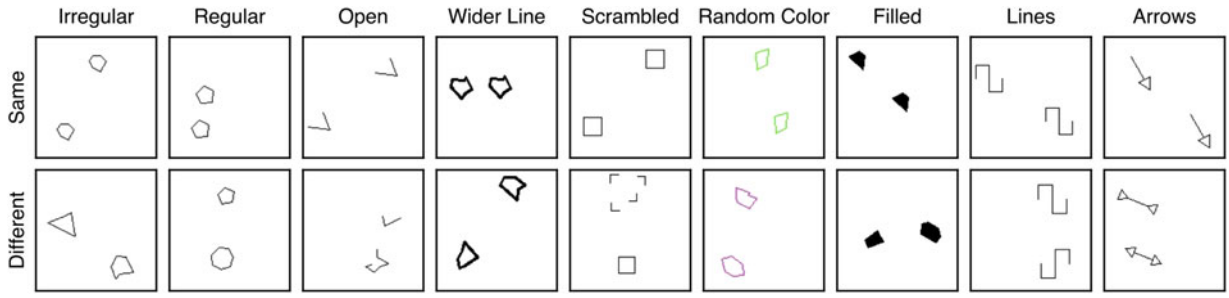


Figure 13. Example stimuli taken from nine different stimulus sets, with the same trials depicted on the top row, different trials on the bottom. The level of similarity between stimulus sets varied, with the greatest overlap between the irregular and regular sets, and little overlap between the irregular set on the one hand and the lines or arrow datasets on the other. Image taken from Puebla and Bowers (2022).

same/different judgments 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.

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, Ludwig, Costa, Malhotra, and Bowers (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 under 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 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, Bowers, Ludwig, Costa, and Malhotra (2022) have 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 under 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, Hill, Santoro, Morcos, and Lillicrap (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 under 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, van Steenkiste, and Schmidhuber (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 shortcuts 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, and Arun (2021). They reported some correspondences between humans and DNNs (described in sect. 4.2.8), but also a series of striking discrepancies. Among 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 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 two-

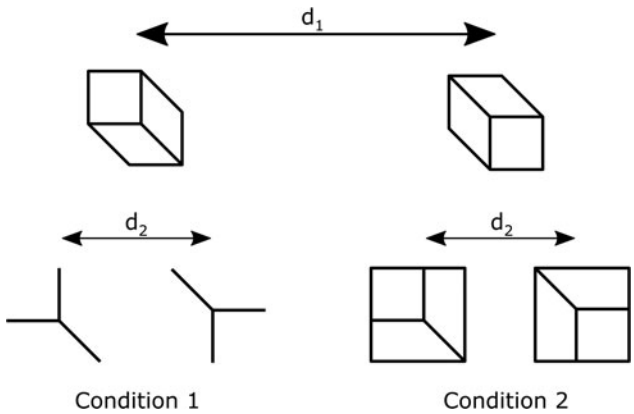


Figure 14. For humans the perceptual distance between the top pair of figures (marked d_1) is larger than the perceptual distance between the two pairs of objects on the bottom (marked d_2). For DNNs, the perceptual distance is the same for all pairs. Images taken from Jacob et al. (2021).

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

dimensional (2D). By contrast, DNNs do not represent the former pair as more dissimilar, suggesting that the models did not improve 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.

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.

4.2.1. Perceptual constancies

Human vision supports a wide range of visual constancies, including color, shape, and lightness constancies, where perceptual judgments 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, Tamuly, Raghunathan, Jain, & Netrapalli, 2020), and consequently, often learn shortcuts (Geirhos et al., 2020a).

4.2.2. Online 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 "online" invariance or tolerance (Blything, Biscione, & Bowers, 2020; Bowers, Vankov, & Ludwig, 2016). Although DNNs can be trained (Biscione & Bowers, 2021; Blything, Biscione, Vankov, Ludwig, & Bowers, 2021) or their architectures modified (Zhang, 2019) to support a range of online 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 among the elements, such that "the whole is more than the sum of its parts." This is not unique to the human cognitive architecture as some nonhuman 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, 2018a; 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.

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,

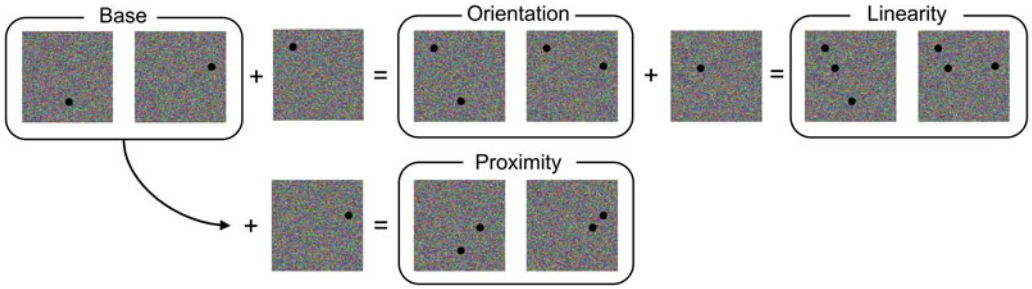


Figure 15. Pomerantz and Portillo (2011) measured Gestalts by constructing a base pair of images (two dots in different locations) and then adding the same context stimulus to each base such that the new image pairs could be distinguished not only using the location of the dots in the base, but also the orientation, linearity, or proximity of the dots. They reported that human participants are faster to distinguish the pair of stimuli under the latter conditions than under the base condition. By contrast, the various DNNs, including DNNs that perform well on Brain-Score, treat the pairs under the orientation, linearity, and proximity conditions as more similar. Images taken from Biscione and Bowers (2023).

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

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 among distractors varies with the visual properties of the target and distractor items (Duncan & Humphreys, 1989; Wolfe, 1994; Wolfe et al., 1989). 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 (Ba et al., 2014; Mnih 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 versus vision for the 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).

4.2.7. Computing shape from nonshape information

Shape is the primary feature that humans rely on when classifying objects, but there are notable examples of recognizing objects based on nonshape 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. (2021)

As discussed above, Jacob et al. (2021) 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 Euclidean distance 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, as noted by Jacob et al., 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

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

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., Adolfi et al., 2023; Feather et al., 2019; Weerts, Rosen, Clopath, & Goodman, 2021). 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 0.02 (or approximately 0.004 of the BOLD variance). In addition, 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 0.02, 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, that is, 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 (sect. 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 nonhuman-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 (e.g., GPT-2 was trained on text taken from 45 million website links and GPT-3 was trained on hundreds 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).

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 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 sect. 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 sect. 4) could be tNote, that the authors of Brain-Score (Schrimpf et al., 2020a, 2020b) have highlighted that

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

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 among 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 categories quickly without catastrophically forgetting previously learned ones (the stability–plasticity dilemma; Grossberg, 1980); among 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 long-standing 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; Holyoak & Hummel, 2000; Marcus, 1998; Pinker & Prince, 1988). 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, 2022; 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 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 regard 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 modeling 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

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

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

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.

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

Competing interest. None.

Notes

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.
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. See http://www.brain-score.org/competition/#workshop.

References

Adolfi, F., Bowers, J. S., & Poeppel, D. (2023). Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Networks*, *162*, 199–211.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W. S., & Nguyen, A. (2019, June). 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*, Long Beach, USA (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*, 1–10.

Baker, N., Kellman, P. J., Erlikhman, G., & Lu, H. (2018a). 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. (2018b). 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*, Stockholm, Sweden (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). Virtual conference.

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. *Journal 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. (2023). Mixed evidence for gestalt grouping in deep neural networks. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-023-00169-2

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*, 1–8.

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., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., & Lerchner, A. (2019). Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 1–22.

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.

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.

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.

Crosby, M., Beyret, B., & Halina, M. (2019). The animal-AI Olympics. *Nature Machine Intelligence*, *1*(5), 257–257.

Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*, *129*(5), 999–1041. 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*, 1–29.

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2022). Some pitfalls of measuring representational similarity using representational similarity analysis. *arXiv preprint*, 1–48. https://www.biorxiv.org/content/10.1101/2022.04.05.487135v1

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2023). Obstacles to inferring mechanistic similarity using Representational Similarity Analysis. *bioRxiv*. https://doi.org/10.1101/2022.04.05.487135

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

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*, 1–33. https://doi.org/10.1101/2022.07.13.499969

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. *Advances in Neural Information Processing Systems*, *32*, 1–12.

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 of the United States of America*, *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. (2020a). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020b). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

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. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

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

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *31*, 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 (New York, N.Y.)*, *358*(6368), eaag2612.

German, J. S., & Jacobs, R. A. (2020). Can machine learning account for human visual object shape similarity judgments. *Vision Research*, *167*, 87–99. https://doi.org/10.1016/j.visres.2019.12.001

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., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., … Lerchner, A. (2019, May). Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, Long Beach, USA (pp. 2424–2433).

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 1–75.

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. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*, 213–227.

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*, 1–23. 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 of the United States of America*, *118*(46), e210477911.

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

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). 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*, 1–32.

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). Erlbaum.

Hummel, J. E. (2013). Object recognition. In D. Reisburg (Ed.), *Oxford handbook of cognitive psychology* (pp. 32–46). Oxford University Press.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517. https://doi.org/10.1037/0033-295X.99.3.480

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

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks 15*(5), 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

Kiat, J. E., Luck, S. J., Beckner, A. G., Hayes, T. R., Pomaranski, K. I., Henderson, J. M., & Oakes, L. M. (2022). Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. *Developmental Science*, *25*, e13155. https://doi.org/10.1111/desc.13155

Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit Gestalt closure. *Computational Brain & Behavior*, *4*, 251–263.

Krauskopf, J. (1963). Effect of retinal image stabilization of the appearance of heterochromatic targets. *Journal of the Optical Society of America*, *53*, 741–744.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

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

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

Kriegeskorte, N., & Wei, X. X. (2021). Neural tuning and representational geometry. *Nature Reviews Neuroscience*, *22*, 703–718. https://doi.org/10.1038/s41583-021-00502-3

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. *Advances in Neural Information Processing Systems*, *32*, 1–12.

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.

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

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, *240*(4853), 740–749.

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. https://doi.org/10.1167/jov.21.10.17

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.

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.

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

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., 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*, 1–51. https://doi.org/10.1101/2021.12.14.472546

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.

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

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

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

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

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing* (Vol. 2). MIT Press.

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

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

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., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J. R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, *35*, 33428–33443.

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*, 1–19.

Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

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*, 1–27. http://arxiv.org/abs/2204.02283

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., & Bowers, J. (2021). The role of disentanglement in generalisation. In *International conference on learning representations*. https://openreview.net/forum?id=qbH974jKUVy

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*, Boston, USA (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. In A. F. Healy & R. W. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (Vol. 4, pp. 177–211). John Wiley & Sons Inc.

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, *144*, 164–175.

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.

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.

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*, 1127–1144.

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. doi:10.1037/A0024330

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*, 179–197.

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. (1988). Perception of shape from shading. *Nature*, *331*(6152), 163–166.

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

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*, Las Vegas, USA (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*, Sydney, Australia (pp. 2940–2949).

Rosenfeld, A., Zemel, R., & Tsotsos, J. K. (2018). The elephant in the room. *arXiv preprint arXiv:1808.03305*, 1–12.

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. *Advances in Neural Information Processing Systems*, *30*, 3856–3866.

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. *Advances in Neural Information Processing Systems*, *35*, 16052–16067.

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*, 1–34.

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 of the United States of America*, *118*(45), e2105646118.

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*, 1–9. 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. *Advances in Neural Information Processing Systems*, *33*, 9573–9585.

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

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, *74*, 1–29.

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

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

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.
Truzzi, A., & Cusack, R. (2020, April). Convolutional neural networks as a model of visual activity in the brain: Greater contribution of architecture than learned weights. Bridging AI and Cognitive Science. In *International conference on learning representations*. https://baicsworkshop.github.io/pdf/BAICS_13.pdf
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. D., & Bowers, J. S. (2023). The role of capacity constraints in convolutional neural networks for learning random versus natural data. *Neural Networks*, *161*, 515–524. https://doi.org/10.1101/2022.03.31.486580
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*, 1–7.
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., & Basri, R. (1991). Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(10), 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, J., Zhang, Z., Xie, C., Zhou, Y., Premachandran, V., Zhu, J., … Yuille, A. (2018). Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications*, *2*(3), 4.
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).
Webb, T. W., Sinha, I., & Cohen, J. D. (2021). Emergent symbols through binding in external memory. *arXiv*, 1–28. https://doi.org/10.48550/arXiv.2012.14601
Weerts, L., Rosen, S., Clopath, C., & Goodman, D. F. (2021). The psychometrics of automatic speech recognition. *bioRxiv, 2021-04*.
Wertheimer, M. (1912). Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, *61*, 161–265 (in German).
Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.
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.
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*, Dublin, Ireland (pp. 957–964).
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, Lille, France (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 of the United States of America*, *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, 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.
Zhang, R. (2019, May). Making convolutional networks shift-invariant again. In *International conference on machine learning*, Long Beach, CA, USA (pp. 7324–7334). Proceedings of Machine Learning Research.
Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*, 3212–3232.
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*, 1–7.
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 of the United States of America*, *118*(3), e2014196118.

Open Peer Commentary

Where do the hypotheses come from? Data-driven learning in science and the brain

Barton L. Anderson[a], Katherine R. Storrs[b] and Roland W. Fleming[c,d]

[a]School of Psychology, University of Sydney, Sydney, Australia; [b]Department of Psychology, University of Auckland, Auckland, New Zealand; [c]Department of Psychology, Justus Liebig University of Giessen, Giessen, Germany and [d]Center for Mind, Brain and Behavior, Universities of Marburg and Giessen, Giessen, Germany
barton.anderson@sydney.edu.au
katherine.storrs@gmail.com
roland.w.fleming@psychol.uni-giessen.de

doi:10.1017/S0140525X23001565, e386

Abstract

Everyone agrees that testing hypotheses is important, but Bowers et al. provide scant details about where hypotheses about perception and brain function should come from. We suggest that the answer lies in considering how information about the outside world could be acquired – that is, learned – over the course of evolution and development. Deep neural networks (DNNs) provide one tool to address this question.

Bowers et al. argue that we need to go beyond hypothesis-blind benchmarking in assessing models of vision. On these points we agree: Benchmarking massively complex models on small and arbitrary brain and behavioural datasets is unlikely to yield satisfying outcomes. The space of models and stimuli is too vast, many models score similarly (e.g., Storrs et al., 2021a), even bad models can score highly in constrained settings, and the approach gives little insight into what model get right or wrong about biological vision. Bowers et al. advocate for the importance of hypothesis-driven research, which is something with which we also agree. However, this is also fraught with challenges for which the authors offer little in the way of solutions. Where do principled hypotheses come from? What theoretical considerations should constrain the hypotheses we consider? We suggest that the different answers that have been proposed to such questions can provide insight into the role that deep neural networks (DNNs) might play in understanding perception and the brain more broadly.

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

One approach is to begin with what our visual systems seem to do and work backwards to "reverse engineer" the brain. This approach was advocated by Marr, and has been taken up by Bayesian approaches that treat visual processes a set of "natural tasks." The idea is that natural selection shaped our brains to approximate "ideal observers" of some set of environmental properties. But natural selection can only act *retrospectively*; it provides no insight into the genesis of the "options" that it "chooses" between. In the case of vision, it could select between brains that were *better* at extracting some world properties than others, but it provides no insight into how brains *discovered* that those environmental properties exist. Something more is needed to explain how the "natural tasks" the brain putatively solves were discovered.

It is here we believe the idea of "data-driven" processes plays a fundamental role. The only known mechanism for getting knowledge about the world into our heads is through our senses. As our brains were not given a list of scene variables they need to estimate, they had to discover properties of the world based on the "diet" of images they experienced over the course of evolution and development. This simple (and seemingly tautological) assertion has a profound theoretical ramification: It implies that anything our brains extract about the world must be based on information *contained in*, or *derivable from*, the input. Two routes to defining "principled hypotheses" about visual function follow from this: (1) we should identify what that information is, that is, explore how what we experience about the world relates to what exists in the input; and (2) we should identify how sensitivity to that information was acquired (learned) over the course of evolution and development, that is, explore the mechanisms that underlie sensitivity to these quantities. These two ideas are not independent. Understanding how information can be learned from images can provide insight into what is learned, and understanding what information is used can constrain attempts to construct a learning process that could become sensitive to it.

To ground this idea, consider an example from our recent work. Psychophysical studies revealed that the subjective perception of surface gloss depends not only on the physical specular reflectance of surfaces, but also on other, physically independent scene properties such as shape and illumination (Ho, Landy, & Maloney, 2008). Further experiments revealed that that these perceptual errors were caused by differences in the spatial structure and distribution of specular reflections in the image, which relates our experience of a world property (gloss) to properties of images (Marlow, Kim, & Anderson, 2012). We then showed that a system trained to recover "ground truth" (i.e., trained to learn a mapping between images and gloss) failed to predict human judgements. In contrast, unsupervised DNNs, designed to summarise and predict properties of the input, learned representations exhibiting the same pattern of successes and errors in perceived gloss as humans (Storrs, Anderson, & Fleming, 2021b). In essence, the unsupervised DNN partially – but imperfectly – disentangled the different scene variables (here, gloss, shape, and illumination) in similar ways as our visual system. DNNs thus provided insight into how such illusions could result from an (imperfect) learning process.

How might we go about discovering the information in images that our visual systems use, and what role might DNNs play in *hypothesis formation* and/or *model development*? Supervised DNNs of the variety Bowers et al. focus on are inherently teleological; they start with a goal, and coerce the system towards that goal through explicit training of the distinctions it wants the system to make (Yamins & DiCarlo, 2016). Unsupervised or self-supervised DNNs are techniques for finding similarities and differences between general features based on statistical properties of the input. One view of such networks is that they are generalised "covariance detectors," which are driven largely by how different image properties do or do not covary. The idea that the visual system derives information about scene properties from the way that different types of image structure covary has provided recent leverage in understanding how the brain extracts the shape and material properties of surfaces (Anderson & Marlow, 2023; Marlow & Anderson, 2021; Marlow, Mooney, & Anderson, 2019).

What role might DNNs take in *evaluating* different computational models of vision? We agree that the psychology and psychophysics literatures provide a wealth of excellent starting points for testing candidate computational models of vision. However, human ingenuity is not always well suited to designing stimuli or experiments that can differentiate between multiple computationally complex alternative models – and such complexity will be unavoidable if we seek to capture human vision in even broad strokes. Therefore, we also see value in using automated selection of complex stimuli to maximally differentiate complex models (e.g., Golan, Raju, & Kriegeskorte, 2020; Wang & Simoncelli, 2008). More broadly, deep learning provides a means to instantiate different hypotheses about how vision is acquired, and the impacts this has on the mature visual system. There is a lot more to deep learning than benchmarking.

Financial support. Funding (DP210102218) for this work was from the Australian Research Council to B. L. A.; by a Marsden Fast Start grant (MFP-UOA2109) from the Royal Society of New Zealand to K. R. S.; and by the DFG (222641018-SFB/TRR-135 TP C1) and Research Cluster "The Adaptive Mind," funded by the Hessian Ministry for Higher Education, Research, Science and the Arts to R. W. F.

Competing interest. None.

References

Anderson, B. L., & Marlow, P. J. (2023). Perceiving the shape and material properties of 3D surfaces. *Trends in Cognitive Sciences*, *27*(1), 98–110. doi:10.1016/j.tics.2022.10.005

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

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*, *19*, 196–204.

Marlow, P., & Anderson, B. (2021). The cospecification of the shape and material properties of light permeable materials. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(14), e2024798118.

Marlow, P., Kim, J., & Anderson, B. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1–5.

Marlow, P., Mooney, S., & Anderson, B. (2019). Photogeometric cues to perceived surface shading. *Current Biology*, *29*(2), 306–311.

Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021b). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behavior*, *5*, 1402–1417. https://doi.org/10.1038/s41562-021-01097-6

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021a). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064.

Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, *8*(12), 8.

Yamins, D., & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365. https://doi.org/10.1038/nn.4244

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

Even deeper problems with neural network models of language

Thomas G. Bever, Noam Chomsky, Sandiway Fong and Massimo Piattelli-Palmarini

Department of Linguistics, University of Arizona, Tucson, AZ, USA
tgb@arizona.edu
nchomsky3@gmail.com
Sandiway@arizona.edu
Massimo@arizona.edu
https://bever.info
https://chomsky.info
https://sandiway.arizona.edu
https://massimo.sbs.arizona.edu

doi:10.1017/S0140525X23001619, e387

Abstract

We recognize today's deep neural network (DNN) models of language behaviors as engineering achievements. However, what we know intuitively and scientifically about language shows that what DNNs are and how they are trained on bare texts, makes them poor models of mind and brain for language organization, as it interacts with infant biology, maturation, experience, unique principles, and natural law.

There is a long tradition of taking current engineering devices as models of how nature itself works. Starting in the eighteenth century, new techniques to control motion in practical uses were adapted to create lifelike models of animals and humans, often populating royal gardens with marvelous lifelike creatures. It was tempting to suggest that the brain itself operates with its own instantiation of such mechanistic principles. Early on came the analogy with a clock, then the telephone network, then the digital computer. Today's "deep neural networks" (DNNs) are sometimes taken as models, since they achieve fantastic performance accuracy in human mental activities, discriminating objects, or producing normal language.

Bowers et al. question the utility of using DNNs and related methodology as models of vision with optical object recognition and categorization as a rubric: They note a range of empirical failures, experimental flaws, and principled reasons why DNNs fail to include other vision facts.

Vision in large part organizes the independent physical world, but human language lies at the internal extreme – almost completely created by the human mind/brain. Accordingly, investigations of language necessarily start with study of the internal knowledge itself. What such investigations reveal about language is inconsistent with DNNs that eschew linguistic theories.

(a) The poverty of the stimulus for children – limited experience, and little explicit training result in sophisticated language ability.
(b) The immediate role in early child language of hierarchical categories and computational constraints (e.g., on anaphoric relations).
(c) Structural elements of language syntax are discrete, the number of combinations is infinite.
(d) The distinction between grammar (aka *competence*) and behavior (aka *performance*) (note: DNNs are intentionally dependent on actual language behaviors).
(e) The role of maximum computational simplicity underlying nature (Einstein's *Miracle Creed*; Chomsky, in press; McDonough, 2022).

Bowers et al. note the notorious flaw of DNNs: "…state-of-the-art DNNs of natural language processing receive training that far exceeds any human experience…. 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)" (target article, sect. 5, para. 2).

The term "inductive biases" reflects an assumption that the "poverty of the stimulus" can be overcome by a list of built-in "priors" which increase the speed of gradual inductive learning: Yet decades of research show that that language emerges without any such general induction process. Rather the evidence indicates an available universal grammar, which defines limited structural options for all languages. Children quickly latch onto particular options of their native language from "signature sentences" (e.g., Gleitman & Landau, 2013; Guasti, 2002; Yang, 2011).

The study of vision and language do share some features. For example, in both domains, the structure is often clarified by subtle cues, illuminating its critical properties. In vision, an image of a panda plus a little visual noise is reported as a gibbon for trained DNNs (Goodfellow et al., 2015). Correspondingly in language research, "minimal pairs" (sentences that vary slightly), can result in strong and reliable differences in structure, interpretation, and grammaticality. Language's discrete infinity property ensures an endless supply of such examples. Thus, large-scale DNN systems, despite unlimited storage, and vast amounts of language data, do not reliably match human performance: Imitation without the human language faculty.

Humans recognize that "the chicken is ready to eat" exhibits structural ambiguity. DNN systems that explicitly compute parses, for example, Google Natural Language, do not recognize the ambiguity, preferring the sentential subject to be subject of "eat." Generative artificial intelligence (AI) systems do not output parses, but we can still deduce underlying grammatical relations by appending a question. In the case of ChatGPT, we can ask for comment with "Is X an ambiguous sentence?" This line of questioning reveals that it assumes "chicken" is the object of "eat." Swapping "children" for "chicken" reveals ambiguity that it reports quite disturbing. Context, for example, the relative proximity of discourses involving cannibals, the story of Hansel and Gretel, or hungry aliens, plays a relative role in ChatGPT's training.

In fact, ChatGPT uses a several thousand token context, potentially capturing discourse phenomena. Consider "The white rabbit jumped from behind the bushes. The animal looked around and then he ran away." For both humans and ChatGPT, *he*, the *animal* and *white rabbit* are preferred to be the same. But if the sentences are reversed in order, only humans then treat "rabbit" as a different entity from "animal," revealing a fundamental principle of anaphoric relations. If DNN is to be a useful model of human behavior, we must know which parameter out of the billions should be adjusted to correct such divergence: Within the statistical enterprise, such errors cannot be diagnosed nor fixed.

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

The authors briefly raise issues involving the evolution of vision as constraining it gradually over many species and eons. Most obvious, and important for vision science, cross-species analogies are multiple and detailed, but not available for language. The authors correctly say that, in spite of claimed success at learning languages, "DNNs will also happily learn [number agreement] in impossible languages with…structures that are not found within any natural languages and which humans struggle to process" (target article, sect. 5, para. 2).

This difference between real syntactic rules and impossible syntactic rules goes much deeper. Like DNNs, humans can master both kinds of rules. Yet in humans, this has underlying neurological correlates that reflect what we know independently about normal neurological processing of language (e.g., Musso et al., 2003). Learning a real language previously unknown to the subjects activates Broca's area: But the same task with an impossible syntactic rule (e.g., a rule that ignores hierarchical structure in favor of serial position) activates only brain areas normally activated during general problem-solving.

We have reviewed ways in which DNNs are empirically inadequate and discordant with theories of language in humans. Adequate or not, we have no idea how individual trained DNNs do what they do: For a DNN to be psychologically useful, we need a theory of the "psychological" innards of the DNN, which is either the same as the theory of human innards, or a unique theory of how initially random associations are compiled from actual behaviors into a model that can be tested on humans (Bever, Fodor, & Garrett, 1968).

Why not focus on attempts to organize and constrain DNNs and other types of models so they comport with what we already know about language, language learning, language representations, and language behaviors? The answer for DNNs is also their touted practical virtue, they learn from actual text, free of hand tailored structural analysis. This engineering virtue pyrrhically underlies why they are doomed to be largely useless models for psychological research on language.

Acknowledgments. We thank Jay Keyser, Andrea Moro, and Robert Berwick for their advice.

Competing interest. None.

References

Bever, T. G., Fodor, J. A., & Garrett, M. (1968). A formal limitation of associationism. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 582–585). Prentice Hall.

Chomsky, N. (in press). The miracle creed and SMT. In M. Greco & D. Mocci (Eds.), *A Cartesian dream: A geometrical account of syntax: In honor of Andrea Moro*. Rivista di Grammatica Generativa/Research in Generative Grammar. Lingbuzz Press.

Gleitman, L., & Landau, B. (2013). Every child an isolate: Nature's experiments in language learning. In M. Piattelli-Palmarini & R. C. Berwick (Eds.), *Rich languages from poor inputs* (pp. 91–106). Oxford University Press.

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Y. Bengio & Y. LeCun (Eds.), *3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9*.

Guasti, M. T. (2002). *Language acquisition: The growth of grammar*. MIT Press.

McDonough, J. K. (2022). *A miracle creed: The principle of optimality in Leibniz's physics and philosophy*. Oxford University Press.

Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Buechel, C., & Weiler, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*(7), 774–781.

Yang, C. (2011). Learnability. In T. Roeper & J. de Villiers (Eds.) *Handbook of language acquisition* (pp. 119–154). Kluwer.

Psychophysics may be the game-changer for deep neural networks (DNNs) to imitate the human vision

Keerthi S. Chandran[a,b], Amrita Mukherjee Paul[a,c], Avijit Paul[d] and Kuntal Ghosh[b]

[a]Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India; [b]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India; [c]Applied Sciences, IIIT Allahabad, Prayagraj, UP, India and [d]Biomedical Engineering, Tufts University, Medford, MA, USA
keerthischandran@gmail.com
rss2020501@iiita.ac.in
avijit.paul@tufts.edu
kuntal@isical.ac.in

doi:10.1017/S0140525X23001759, e388

Abstract

Psychologically faithful deep neural networks (DNNs) could be constructed by training with psychophysics data. Moreover, conventional DNNs are mostly monocular vision based, whereas the human brain relies mainly on binocular vision. DNNs developed as smaller vision agent networks associated with fundamental and less intelligent visual activities, can be combined to simulate more intelligent visual activities done by the biological brain.

In keeping with what Turing proposed for the imitation game (Turing, 1950), a good brain-computational model (Kriegeskorte & Douglas, 2018) would not be the one that performs a particular task with equal or greater accuracy than a human being, but rather the one which would be indistinguishable from a human being vis-à-vis input and output. Psychophysics, interestingly, is also about input and output with the brain as black-box in between (Read, 2015). Bowers et al. provide a comprehensive presentation of the incongruence between deep neural networks (DNNs) and the visual brain, but fails to note this relevant connection of psychophysics to neuroscience for brain-computational modeling (Read, 2015).

Psychophysics is "the analysis of perceptual processes by studying the effect on a subject's experience or behavior of systematically varying the properties of a stimulus along one or more physical dimensions" (Bruce, Green, & Georgeson, 2003). The psychophysics stimulus for vision can be an image or video, and DNN, an information-processing system, may model the subject's response to the stimulus using supervised learning. David Marr had proposed that an information processing system should be understood at three levels: computational, algorithmic, and implementation. The psychophysics task describes the computational level problem, a DNN that performs the same task in silica would represent the algorithmic level, and the electrophysiological or fMRI data obtained during the task will be a by-product of the implementation of the algorithm in the biological brain. If the DNN is considered for an equivalent mapping between input and output as in a psychophysics experiment, then the inputs can be represented by a tensor, whether it is an image, video, sound signal, or a spatially invariant visual stimulus like the flicker; the output would also have a numerical representation which, in case of psychophysics experiments, could be some

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

classification, perceived brightness, color, shape, size, motion, intensity at a particular location in the input signal, or a comparison between two of those perceived sensations at different locations of the stimulus, separated by space or time or both. The algorithm used to transform the stimulus input to output will not be evident from psychophysics experiments, but DNNs can construct that algorithm without its exact knowledge for the programmer.

The dataset can be prepared by manipulating physical parameters associated with the stimulus and getting the subject response for each of the stimuli. There can be some subjective differences between the psychophysics data of human subjects for the same stimuli (Read, 2015). So, it will be a better strategy to train and test a DNN on the psychophysics data of the same subject. Kubota, Hiyama, and Inami (2021) have used psychophysics data obtained from brightness illusions to train DNNs. Kubota et al. (2021) have shown that it is possible to make comparisons between human perception on the one hand, and the output with the said methodology, on the other. DNNs may also be tested on a stimulus, completely different from the one it was trained on, if its output layer is of similar representation to that of the new stimulus. Recently, Ghosh and Chandran (2021) proposed such a technique for flicker stimulus. The intermediate outputs of a DNN can be compared with the brain electrophysiological signals as done by Zipser and Andersen (1988), and more recently by Chandran and Ghosh (2021, 2022) with EEG. We argue that more testable models can be constructed by training on less computationally intensive tasks than tasks like object classification into thousands of classes. For instance, a convolutional neural network (CNN) trained for low-level visual tasks gets deceived by brightness and color illusions (Gomez-Villa, Martín, Vazquez-Corral, Bertalmío, & Malo, 2020). DNNs have also been put forth to solve tasks used in experimental psychology like Raven's progressive matrices (Jahrens & Martinetz, 2020). New network models, different from the engineering goal-oriented image classification DNNs, could be constructed for the purpose as was previously done for finding head-centered coordinates of external objects by monkey brain by Zipser and Andersen (1988). It could be easier to make correlations between outputs of intermediate layers of a neural network with fewer neurons and layers with brain signals than complex networks.

Bowers et al. mentions that DNNs trained on ImageNet do not encode three-dimensional (3D) features of objects or their depth as opposed to human vision. The abovementioned DNNs are trained with datasets prepared from cameras with monocular vision. But the mammalian brain gets information from the two eyes and it is known that human subjects with one eye are not so efficient with depth perception (Westlake, 2001). Robots with stereo cameras making use of DNNs are able to do tasks like calculating position of detected fruit from stereo cameras (Onishi et al., 2019). Stereo vision can enable autonomous driving vehicles to do tasks like object detection, 3D information acquisition, and depth perception (Fan, Wang, Junaid Bocus, & Pitas, 2023). The mammalian brain had input from two eyes throughout the course of its evolutionary history. So training DNNs using stereo camera data might be needed to develop the equivalents of many circuits in the brain.

To conclude, psychophysics with DNNs could be used to construct many of the smaller agents that compose the human mind as proposed by Minsky (1988). Vision agents that compose the mind need to be likewise constructed via DNNs, which may be associated with fundamental activities like brightness perception, motion detection, depth perception, or even less intelligent activities than that, in the parallel visual pathways. Neural networks for more complex tasks can be built with a combination of smaller DNNs using shared layers, or by using output from some layers of a DNN as input for layers of another DNN.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Bruce, V., Green, P. R., & Georgeson, M. A. (2003). *Visual perception: Physiology, psychology, & ecology*. Psychology Press.

Chandran, K. S., & Ghosh, K. (2021). Recurrent convolutional neural networks trained by psychophysics data can predict EEG response to flicker. *Perception*, *50*(ECVP2021 Supplement), 1–244. https://doi.org/10.1177/03010066211059887

Chandran, K. S., & Ghosh, K. (2022). An in-silica computation of alpha oscillations from apparently unrelated psychophysics data. https://doi.org/10.21203/rs.3.rs-1862596/v1

Fan, R., Wang, L., Junaid Bocus, M., & Pitas, I. (2023). Computer stereo vision for autonomous driving: Theory and algorithms. *Studies in Computational Intelligence*, 41–70. https://doi.org/10.1007/978-3-031-18735-3_3

Ghosh, K., & Chandran, K. S. (2021). A low-cost device and technique for generating big data in visual psychophysics to train brain models. *Perception*, *50*(ECVP2021 Supplement), 1–244. https://doi.org/10.1177/03010066211059887

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive CNNs for low-level vision tasks: Analysis and implications. *Vision Research*, *176*, 156–174. https://doi.org/10.1016/j.visres.2020.07.010

Jahrens, M., & Martinetz, T. (2020). Solving Raven's progressive matrices with multi-layer relation networks. In *2020 International joint conference on neural networks (IJCNN)*. Jointly organized by the IEEE Computational Intelligence Society (CIS) and the International Neural Network Society (INNS), Glasgow, UK (pp. 1-6). https://doi.org/10.1109/ijcnn48605.2020.9207319

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Kubota, Y., Hiyama, A., & Inami, M. (2021). A machine learning model perceiving brightness optical illusions: Quantitative evaluation with psychophysical data. In *Proceedings of the Augmented Humans International Conference 2021 (AHs '21)*. Association for Computing Machinery, New York, NY, USA (pp. 174–182). https://doi.org/10.1145/3458709.3458952

Minsky, M. (1988). Prologue. In *The society of mind* (p. 17). Simon & Schuster.

Onishi, Y., Yoshida, T., Kurita, H., Fukao, T., Arihara, H., & Iwai, A. (2019). An automated fruit harvesting robot by using deep learning. *ROBOMECH Journal*, *6*(1), 13. https://doi.org/10.1186/s40648-019-0141-2

Read, J. C. A. (2015). The place of human psychophysics in modern neuroscience. *Neuroscience*, *296*, 116–129. https://doi.org/10.1016/j.neuroscience.2014.05.036

Turing, A. M. (1950). I. – Computing machinery and intelligence. *Mind; A Quarterly Review of Psychology and Philosophy*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/lix.236.433

Westlake, W. (2001). Is a one eyed racing driver safe to compete? Formula one (eye) or two? *British Journal of Ophthalmology*, *85*(5), 619–624. https://doi.org/10.1136/bjo.85.5.619

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679–684. https://doi.org/10.1038/331679a0

A deep new look at color

Jelmer Philip de Vries[a], Alban Flachot[b], Takuma Morimoto[a,c] and Karl R. Gegenfurtner[a]

[a]Department of Psychology, Justus Liebig Universitat, Giessen, Germany; [b]Department of Psychology, York University, Toronto, ON, Canada and [c]Department of Experimental Psychology, University of Oxford, Oxford, UK
vriesdejelmer@gmail.com; flachot.alban@gmail.com; takuma.morimoto@psy.ox.ac.uk; gegenfurtner@uni-giessen.de
www.jelmerdevries.com; https://www.allpsych.uni-giessen.de/karl/; https://sites.google.com/view/tmorimoto

doi:10.1017/S0140525X23001620, e389

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

Abstract

Bowers et al. counter deep neural networks (DNNs) as good models of human visual perception. From our color perspective we feel their view is based on three misconceptions: A misrepresentation of the state-of-the-art of color perception; the type of model required to move the field forward; and the attribution of shortcomings to DNN research that are already being resolved.

One of the main arguments that Bowers et al. put forth is that deep neural networks (DNNs) classify objects in a fundamentally different manner from humans. However, what Bowers et al. promote as the state-of-the-art in terms of color processing, namely a strict segregation of visual streams for color and shape (Livingstone & Hubel, 1987), is outdated and has repeatedly been rejected (see Garg, Li, Rashid, & Callaway, 2019; Gegenfurtner & Kiper, 2003; Shapley & Hawken, 2011). The fact that line drawings can be recognized quickly does not imply that object processing in humans does not rely on color. On the contrary, boundaries defined by color appear essential for image segmentation in humans (Hansen & Gegenfurtner, 2009, 2017; Shapley & Hawken, 2011). Moreover, the view on how color is represented in the brain has evolved from one of having a single-color center, to one where color-biased regions are found throughout the ventral stream (e.g., Conway, 2018; Gegenfurtner, 2003). While classical algebraic models of color vision have been highly successful in explaining the processing in the cones and color-opponent stages in the eye, higher level cortical processing is still not well understood (for a recent review, see Siuda-Krzywicka & Bartolomeo, 2020). All the evidence points toward an integral role for color in extracting objects, and this perfectly matches the emphasis that DNNs place on objects rather than isolated features.

Bowers et al. emphasize the lack of experimental rigor in testing DNNs compared to testing humans. While we largely agree, it is also important to consider the limitations of a myopic drive to constrain experiments to single-feature manipulations. Reductionist experimental methodology in human research typically diverges greatly from our natural experiences. It biases research toward investigating cognitive functions not necessarily at the core of how our system operates in daily life (e.g., Shamay-Tsoory & Mendelsohn, 2019). While Biederman's (1987) geons may be sufficient for recognizing isolated objects in stereotypical configurations, in daily scenes objects appear in countless varying states (e.g., a cat curled up in a ball on the couch) and other features will gain importance. To avoid a reductionist bias, neuroscientific models need to be grounded in behavior natural to the organism (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). DNNs may seem misplaced within a classical approach of model-based hypothesis testing, where strongly reductive process models are defined explicitly to test a single hypothesis. However, DNNs are highly suitable for studying how behavior shapes underlying mechanisms. By observing emerging properties in the context of learning specific tasks and manipulating input we can develop improved hypotheses on *why* colors are represented the way that they are in the human visual system. Subsequently, as DNNs allow for comparisons at many levels, from single trials in psychophysics or electrophysiology experiments, up to derived mental representations in a human observer, they make it possible to study *how* these mechanisms may be implemented.

Naturally, without considerable overlap with the human visual system we would not consider DNNs adequate models of human vision. However, Bowers et al. focus strongly on discrepancies between humans and DNNs, but neglect important overlap. For color, properties of artificial neurons show great overlap with those in primate visual cortex: Many neurons exhibit double-opponent receptive fields (Flachot & Gegenfurtner, 2018, 2021; Rafegas, Vanrell, Alexandre, & Arias, 2020; Rafegas & Vanrell, 2018), and a moderate functional segregation between color and achromatic information was found at the early stages, corresponding to retino-geniculate processing (Flachot & Gegenfurtner, 2018). On a higher level, DNNs were also shown to outperform classical models in identifying regions of the objects that are highly predictive of human behavioral patterns when discriminating color of naturalistic objects (Ponting, Morimoto, & Smithson, 2023). Moreover, qualitative similarities have been uncovered between DNNs and human participants in color constancy experiments where individual cues known to affect human color constancy were manipulated (Flachot et al., 2022). Finally, in our efforts to uncover *why* humans adopt a categorical representation of color, we found that DNNs trained specifically for object recognition incorporate a categorical representation of color that is highly similar to that of humans (de Vries, Akbarinia, Flachot, & Gegenfurtner, 2022).

In that study, we strongly focused on translating psychophysical methods to DNNs. We created a match-to-sample task inspired by work on color categorization in pigeons (Wright & Cumming, 1971) using controlled-stimuli and, in secondary experiments, validated their use in the DNN. We also translated the concept of categorical color perception (where colors from different categories are distinguished faster than those from the same category) to the DNN. Our study on color constancy, mentioned above, also included the typical manipulations found in psychophysical studies on the issue. Finally, the studies on neural color tuning translated methods from single-cell recordings in nonhuman primates to DNNs. Together, this introduces important tools to move beyond purely correlational human–DNN comparisons and to investigate where the DNN is similar to the human visual system and where it deviates. Carefully designed experiments allow for collecting response patterns from DNNs through which richer human–DNN comparisons are possible. Notably, our findings are not purely correlational in nature. For example, to establish whether object recognition was important to finding a categorical representation of color, we contrasted an object-trained DNN with the one trained to distinguish artificial from man-made scenes. Importantly, human-like color categories were only found for the object task, indicating that object learning may be crucial in shaping human color categories.

Color and object processing are intricately connected (e.g., Conway, 2018; Witzel & Gegenfurtner, 2018) and understanding color perception will require a model that takes objects into account. DNNs enable us to investigate under what circumstances color phenomena arise and to inspect how they are implemented. Naturally, shortcomings, such as a reliance on correlation-based comparisons and strong divergences from human object processing should be addressed. However, we believe these shortcomings will prove predominantly temporary in nature, as they are already being taken into account in several recent studies. As such, where

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

Bowers et al. take issue with using object-trained DNNs, we see opportunity.

Financial support. This study is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 222641018 – SFB/TRR 135 TP C2, and European Research Council Advanced Grant Color 3.0 (884116). A. F. is funded by a VISTA postdoctoral fellowship. T. M. is supported by a Sir Henry Wellcome Postdoctoral Fellowship from Wellcome Trust and a Junior Research Fellowship from Pembroke College, University of Oxford.

Competing interest. None.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.4324/9781351156288-24

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual Review of Vision Science*, *4*(1), 381–402. https://doi.org/10.1146/annurev-vision-091517-034202

de Vries, J. P., Akbarinia, A., Flachot, A., & Gegenfurtner, K. R. (2022). Emergent color categorization in a neural network trained for object recognition. *eLife*, *11*, e76472. https://doi.org/10.7554/eLife.76472

Flachot, A., Akbarinia, A., Schütt, H. H., Fleming, R. W., Wichmann, F. A., & Gegenfurtner, K. R. (2022). Deep neural models for color classification and color constancy. *Journal of Vision*, *22*(4), 1–24. https://doi.org/10.1167/jov.22.4.17

Flachot, A., & Gegenfurtner, K. R. (2018). Processing of chromatic information in a deep convolutional neural network. *Journal of the Optical Society of America A*, *35*(4), B334. https://doi.org/10.1364/josaa.35.00b334

Flachot, A., & Gegenfurtner, K. R. (2021). Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vision Research*, *182*, 89–100. https://doi.org/10.1016/j.visres.2020.09.010

Garg, A. K., Li, P., Rashid, M. S., & Callaway, E. M. (2019). Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science (New York, N.Y.)*, *364*(6447), 1275–1279. https://doi.org/10.1126/science.aaw5868

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, *4*(7), 563–572. https://doi.org/10.1038/nrn1138

Gegenfurtner, K. R., & Kiper, D. C. (2003). *Annual review of neuroscience*, *26*, 181–206. https://doi.org/10.1146/annurev.neuro.26.041002.131116

Hansen, T., & Gegenfurtner, K. R. (2009). Independence of color and luminance edges in natural scenes. *Visual Neuroscience*, *26*(1), 35–49. https://doi.org/10.1017/S0952523808080796

Hansen, T., & Gegenfurtner, K. R. (2017). Color contributes to object-contour perception in natural scenes. *Journal of Vision*, *17*(3), 1–19. https://doi.org/10.1167/17.3.14

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, *7*(11), 3416–3468. https://doi.org/10.1523/jneurosci.07-11-03416.1987

Ponting, S., Morimoto, T., & Smithson, H. (2023). Modelling surface color discrimination under different lighting environments using image chromatic statistics and convolutional neural networks. *Journal of the Optical Society of America A*, *40*(3), A149–A159. https://doi.org/10.1364/josaa.479986

Rafegas, I., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision Research*, *151*, 7–17. https://doi.org/10.1016/j.visres.2018.03.010

Rafegas, I., Vanrell, M., Alexandre, L. A., & Arias, G. (2020). Understanding trained CNNs by indexing neuron selectivity. *Pattern Recognition Letters*, *136*, 318–325.

Shamay-Tsoory, S. G., & Mendelsohn, A. (2019). Real-life neuroscience: An ecological approach to brain and behavior research. *Perspectives on Psychological Science*, *14*(5), 841–859. https://doi.org/10.1177/1745691619856350

Shapley, R., & Hawken, M. J. (2011). Color in the cortex: Single- and double-opponent cells. *Vision Research*, *51*(7), 701–717. https://doi.org/10.1016/j.visres.2011.02.012

Siuda-Krzywicka, K., & Bartolomeo, P. (2020). What cognitive neurology teaches us about our experience of color. *Neuroscientist*, *26*(3), 252–265. https://doi.org/10.1177/1073858419882621

Witzel, C., & Gegenfurtner, K. R. (2018). Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, *4*, 475–499.

Wright, A. A., & Cumming, W. W. (1971). Color-naming functions for the pigeon. *Journal of the Experimental Analysis of Behavior*, *15*(1), 7–17.

Let's move forward: Image-computable models and a common model evaluation scheme are prerequisites for a scientific understanding of human vision

James J. DiCarlo[a], Daniel L. K. Yamins[b], Michael E. Ferguson[a], Evelina Fedorenko[a], Matthias Bethge[c], Tyler Bonnen[b] and Martin Schrimpf[a,d]

[a]Dept. of Brain and Cognitive Sciences, Quest for Intelligence, and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA; [b]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA; [c]Tübingen AI Center, University of Tübingen, Tübingen, Germany and [d]École polytechnique fédérale de Lausanne, Lausanne, Switzerland
dicarlo@mit.edu; https://dicarlolab.mit.edu
yamins@stanford.edu
mferg@mit.edu
evelina9@mit.edu; https://evlab.mit.edu/
matthias.bethge@bethgelab.org; https://bethgelab.org/
bonnen@stanford.edu; http://neuroailab.stanford.edu/research.html
msch@mit.edu; https://mschrimpf.com/

doi:10.1017/S0140525X23001607, e390

Abstract

In the target article, Bowers et al. dispute deep artificial neural network (ANN) models as the currently leading models of human vision without producing alternatives. They eschew the use of public benchmarking platforms to compare vision models with the brain and behavior, and they advocate for a fragmented, phenomenon-specific modeling approach. These are unconstructive to scientific progress. We outline how the Brain-Score community is moving forward to add new model-to-human comparisons to its community-transparent suite of benchmarks.

Common ground

As vision scientists, we believe that an understanding of human visual processing should ultimately explain all visually driven behavior. Because vision operates – by definition – on visual input, a science of human vision ultimately requires "image-computable" models and theories that produce those models. Bowers et al. endorse this view as every psychology experiment they suggest focuses on the effects of manipulations of combinations of image pixels.

On empirical tests of vision models

As empirical vision scientists, we also believe that advances in understanding visual processing will arise from rigorous, community-transparent tests of model predictions against empirical observations from the brain (e.g., patterns of neural firing) and the mind (e.g., patterns of behavior). As such, we and others have contributed to the creation of an open-source platform where any member of the vision community can find the leading models, test new models, see the most model-disruptive

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

experimental benchmarks, and add new benchmarks (www.brain-score.org; Schrimpf et al., 2018, 2020).

The most constructive contribution of Bowers et al. is the identification of a set of human behavioral vision findings that the authors believe will not be well-predicted by currently leading deep artificial neural network (ANN) models (target article, sect. 4.1). To evaluate this claim, the Brain-Score community is turning these empirical findings into accessible benchmarks that current (and future) models of human visual processing can be evaluated on. The results of this evaluation, especially if these benchmarks indeed present a challenge for current ANN models, should and would motivate next steps in human vision modeling. We report the following status at the time of this writing:

- We have implemented a benchmark based on Baker and Elder (2022). We find that some ANN vision models are within the noise ceiling of the human data (based on resampling of the human data).
- Two of the papers (Puebla & Bowers, 2022; Zhang, Bengio, Hardt, Recht, & Vinyals, 2021) evaluate the performance of some ANN models without a human reference. Thus these studies currently provide no empirical support for the target article's claim that current ANN models fail to capture human behavior. But human data could be collected to turn these into benchmarks.
- Three of the papers (Bowers & Jones, 2007; Mack, Gauthier, Sadr, & Palmeri, 2008; Saarela, Sayim, Westheimer, & Herzog, 2009) produced human behavioral data that ANN models do not yet have a standardized way to make predictions about, for example, reaction times. This is surmountable (e.g., Spoerer, Kietzmann, Mehrer, Charest, & Kriegeskorte, 2020) and we view this as a goal for future models and for Brain-Score.

On current vision models

We are not dogmatically committed to any current deep ANN model of human vision, none of which are perfect models of human vision, as the Brain-Score effort helped illuminate. However, we disagree with Bowers et al.'s claim that deep ANNs are not the currently leading models of human ventral visual processing. Bowers et al. critique ANN models without offering a better alternative: They imply that better models exist or should exist, but do not elaborate on what those models are. In the absence of an alternative model, it is justifiable to refer to ANNs as the currently best models. In fact, as can be seen on Brain-Score, in addition to the ability of some ANN models to moderately well predict neural responses at multiple visual processing stages, those same ANN models do, to some extent, predict even quite challenging behavioral data patterns (Geirhos et al., 2021; Rajalingham et al., 2018).

Bowers et al. eschew community-transparent suites of benchmarks yet they imply an alternative notion of vision model evaluation, which is somehow not a suite of benchmarks. But again, they do not produce a feasible alternative. Of course, the model rankings produced by benchmarks also depend on the choice of datasets and metric used for evaluation. We will continue to help the Brain-Score community expand the range of datasets and we are not dogmatically committed to any particular choice of metric. Different subcommunities may prefer to initially focus on different metrics (e.g., to know the currently best behavioral model regardless of underlying brain alignment, or vice-versa), and Brain-Score should support those different benchmark weightings. But we see no alternative to support advances in models of vision other than an open, transparent, and community-driven way of model comparison.

On building new vision models

Bowers et al. appear to favor a classic approach in which a separate model is built for each psychological phenomenon, using specialized stimuli that are hand-crafted to enable certain visual features to be well-defined – for example, illusory contours or shape primitives. The appeal of this approach is that it reduces the complexity of a high-dimensional pixel input space into small intuitive sets of features that enable the formulation and testing of conceptual hypotheses about vision – for example, the mechanisms of a particular class of visual illusions. However, because this approach requires dramatically restricting the stimuli under consideration, such hypotheses often cover a near-zero fraction of image space. In our opinion, the idea that a universal scientific model of human vision will result from sets of fragmented explanations that only engage a tiny fraction of image space is illusory (Newell, 1973).

In contrast, the approach of starting with image-computable models that we favor enables tangible progress toward a unified model of human vision. Transparent tracking of model shortcomings lights the path to this goal. We acknowledge that the image-computability requirement may make formulation of traditional conceptual tests of a model more challenging. But it, by no means, makes such tests impossible. Any pattern of behavioral data, including those discussed in the target article, should be translatable into a behavioral benchmark on Brain-Score.

Moving forward

Ultimately, we think that the advantages that image-computable models have in enabling evaluation of predictions about diverse visual stimuli and phenomena heavily outweighs their disadvantages. And maintaining and expanding a common evaluation scheme for image-computable models of vision is, in our view, a prerequisite for channeling the valuable contributions of vision science – across neuroscience, cognitive science, psychology, and computer vision – toward convergence on the best scientific models of human vision. Let's move forward!

Acknowledgments. We thank Kohitij Kar, Micheal Lee, Nancy Kanwisher, Nikolaus Kriegeskorte, and Chris Shay for helpful discussions and support.

Financial support. This work was supported in part by the Semiconductor Research Company (SRC) and DARPA (J. J. D.), Simons Foundation (542965, J. J. D.), Office of Naval Research (MURI N00014-21-1-2801; N00014-20-1-2589, J. J. D., D. L. K. Y.), and National Science Foundation (2124136, J. J. D.).

Competing interest. M. B. is a co-founder of Maddox AI. All other authors have no competing interest.

References

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

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

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. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

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

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.
Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. Visual information processing. Academic Press.
Puebla, G., & Bowers, J. S. (2022). Can deep convolutional neural networks support relational reasoning in the same-different task? *Journal of Vision*, *22*(10), 1–18.
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.
Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, *9*(2), 5.
Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., … DiCarlo, J. J. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007.
Schrimpf, M., Kubilius, J., Lee, M. J., Murty, R., Apurva, N., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423.
Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Computational Biology*, *16*(10), e1008215.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

Implications of capacity-limited, generative models for human vision

Joseph Scott German[a] and Robert A. Jacobs[b]

[a]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA and [b]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA
jgerman@ucsd.edu
rjacobs@ur.rochester.edu
https://www2.bcs.rochester.edu/sites/jacobslab/people.html

doi:10.1017/S0140525X23001772, e391

Abstract

Although discriminative deep neural networks are currently dominant in cognitive modeling, we suggest that capacity-limited, generative models are a promising avenue for future work. Generative models tend to learn both local and global features of stimuli and, when properly constrained, can learn componential representations and response biases found in people's behaviors.

The target article offers cogent criticisms of deep neural networks (DNNs) as models of human cognition. Although discriminative DNNs are currently dominant in cognitive modeling, other approaches are needed if we are to achieve a satisfactory understanding of human cognition. We suggest that generative models are a promising avenue for future work, particularly capacity-limited, generative models designed around componential representations (e.g., part-based representations of visual objects and scenes).

A *generative model* is a model that learns a joint distribution of visible (i.e., observed) and hidden (latent) variables. Importantly, many generative models allow us to sample from the distribution learned by the model, producing "synthetic" examples of the concept modeled by the distribution. Using a generative model to make inferences about external stimuli is a matter of identifying the properties of the generative model most likely to have produced these stimuli. By their very nature, generative models neatly sidestep many of the issues with discriminative models, as described in the target article.

Most obviously yet perhaps most importantly, they are typically judged not based on predictive performance, but on their ability to synthesize examples of concepts, which requires a more profound understanding of those concepts than does mere discrimination, potentially leading to task-general representations capable of explaining far more of human perceptual and cognitive reasoning. For example, unlike discriminative models trained to categorize images, which tend to base their decisions on texture patches and local shape instead of global shape as humans do, a successful generative model must include an understanding of global object shape, as otherwise its samples would not be realistic. Inference in such a generative model would therefore be sensitive to object shape as a matter of course, as well as a number of other properties that might be ignored by a discriminatively trained model.

Another important feature of human cognition not captured by large DNNs is capacity limits. People cannot remember all aspects of a visual environment, and so human vision needs to be selective and efficient. By contrast, DNNs often contain billions of adaptable parameters, providing them with enormous learning, representational, and processing capacities. These seemingly unlimited capacities are in stark contrast to the dramatically limited capacities of biological vision, as noted in the target article. This need for efficiency underlies people's attentional and memory biases. People are biased toward "filling-in" missing features (i.e., features not attended or remembered) with values that are highly frequent in the environment. In addition, people are biased toward attending to and remembering those features which are most relevant for their current goal, thereby maximizing task performance.

Bates, Lerch, Sims, and Jacobs (2019) experimentally evaluated these biases using an optimal model of capacity-limited visual working memory (VWM) based on "rate-distortion theory" (RDT; see Sims, Jacobs, & Knill, 2012). Both biases were predicted by the RDT model: An optimal VWM should be biased toward allocating its limited memory resources toward high-probability feature values and toward task-relevant features. Bates and Jacobs (2021) studied people's responses in the domain of visual search and attention. The RDT model predicted important aspects of these responses, including "set-size" effects indicative of limited capacity, aspects not accounted for by a model based on Bayesian decision theory.

In accord with these ideas, a popular form of generative model, a "variational autoencoder" (VAE) uses a loss function during training that penalizes a large growth in capacity. A VAE maps an input through one or more hidden layers, with a penalized capacity at one of the layers, to an output layer that attempts to reconstruct the input. Reconstructions are typically imperfect due to the "lossy" representations at the "bottleneck" hidden layer with restricted capacity. Machine learning researchers have shown important mathematical relationships between VAEs and RDT (Alemi et al., 2017, 2018; Ballé, Laparra, & Simoncelli, 2016; Burgess et al., 2018). Bates and Jacobs (2020) used VAEs to model biases and set-size effects in human visual perception and memory. We believe this is an encouraging early step toward developing capacity-limited, generative models of human vision.

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

The desire for efficient representations also leads to componential or part-based approaches, and generative models naturally lend themselves to understanding concepts based on parts and relationships between them, as humans do (in contrast to DNNs, as the target article points out, citing German and Jacobs, 2020, and Erdogan and Jacobs, 2017). The same basic parts can be used to create a wide variety of distinct objects, just by changing the relationships between them (the basis of many perceptual and cognitive models such as Biederman, 1987). Learning new object concepts thereby becomes more efficient, as once a part has been learned, it can be used in the representation and construction of any object concept using it, including new ones. This idea can be further extended by supposing that parts are made out of subparts, and so on, producing hierarchical, componential generative models (e.g., Lake, Salakhutdinov, & Tenenbaum, 2015; Nash & Williams, 2017).

To be sure, a capacity-limited, generative approach is not going to "solve" cognitive modeling overnight. It still faces major obstacles such as computationally expensive inference and a lack of objective criteria with which to judge the quality of its synthesized instances. However, we are optimistic that these issues can be resolved, and we hope the target article inspires researchers to look beyond the established discriminative DNN paradigm. Perhaps if capacity-limited, generative models receive as much research attention and development as discriminative models have, we can look forward to significant advances in both computational cognitive modeling and machine learning.

Financial support. This work was funded by NSF research grants BCS-1824737 and DRL-1561335.

Competing interest. None.

References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). An information-theoretic analysis of deep latent variable models. *arXiv preprint* arXiv:1711.00464. Retrieved from https://arxiv.org/pdf/1711.00464v1.pdf

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018). Fixing a broken ELBO. *arXiv preprint* arXiv:1711.00464v3. Retrieved from https://arxiv.org/pdf/1711.00464v3.pdf

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint* arXiv:1611.01704. Retrieved from https://arxiv.org/pdf/1611.01704.pdf

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*, 891–917.

Bates, C. J., & Jacobs, R. A. (2021). Optimal attentional allocation in the presence of capacity constraints in uncued and cued visual search. *Journal of Vision*, *21*(5), 3, 1–23.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11, 1–23.

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

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., … Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint* arXiv:1804.03599. Retrieved from https://arxiv.org/pdf/1804.03599.pdf

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

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

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science (New York, N.Y.)*, *350*(6266), 1332–1338.

Nash, C., & Williams, C. K. I. (2017). The shape variational autoencoder: A deep generative model of part-segmented 3D objects. *Eurographics Symposium on Geometry Processing*, *36*(5), 1–11.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*, 807–830.

Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses

Tal Golan[a], JohnMark Taylor[b], Heiko Schütt[b,c], Benjamin Peters[d], Rowan P. Sommers[e], Katja Seeliger[f], Adrien Doerig[g], Paul Linton[b,h,i], Talia Konkle[j], Marcel van Gerven[k], Konrad Kording[l,m], Blake Richards[m,n,o,p,q], Tim C. Kietzmann[g], Grace W. Lindsay[r] and Nikolaus Kriegeskorte[b,s]

[a]Department of Cognitive and Brain Sciences, Ben-Gurion University of the Negev, Be'er Sheva, Israel; [b]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA; [c]Center for Neural Science, New York University, New York, NY, USA; [d]School of Psychology & Neuroscience, University of Glasgow, Glasgow, UK; [e]Department of Neurobiology of Language, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; [f]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany; [g]Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany; [h]Presidential Scholars in Society and Neuroscience, Center for Science and Society, Columbia University, New York, NY, USA; [i]Italian Academy for Advanced Studies in America, Columbia University, New York, NY, USA; [j]Department of Psychology and Center for Brain Sciences, Harvard University, Cambridge, MA, USA; [k]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands; [l]Departments of Bioengineering and Neuroscience, University of Pennsylvania, Philadelphia, PA, USA; [m]Learning in Machines and Brains Program, CIFAR, Toronto, ON, Canada; [n]Mila, Montreal, QC, Canada; [o]School of Computer Science, McGill University, Montreal, QC, Canada; [p]Department of Neurology & Neurosurgery, McGill University, Montreal, QC, Canada; [q]Montreal Neurological Institute, Montreal, QC, Canada; [r]Department of Psychology and Center for Data Science, New York University, New York, NY, USA and [s]Departments of Psychology, Neuroscience, and Electrical Engineering, Columbia University, New York, NY, USA

golan.neuro@bgu.ac.il
jt3295@columbia.edu
hs3110@columbia.edu
benjamin.peters@posteo.de
rowan.sommers@mpi.nl
katjaseeliger@posteo.de
adoerig@uni-osnabrueck.de
paul.linton@columbia.edu
talia_konkle@harvard.edu
koerding@gmail.com
blake.richards@mila.quebec
tim.kietzmann@uni-osnabrueck.de
grace.lindsay@nyu.edu
nk2765@columbia.edu
brainsandmachines.org
johnmarktaylor.com
hebartlab.com
kietzmannlab.org
https://linton.vision/
https://konklab.fas.harvard.edu/
artcogsys.com
kordinglab.com
linclab.org
kietzmannlab.org
lindsay-lab.github.io

doi:10.1017/S0140525X23001553, e392

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

Abstract

An ideal vision model accounts for behavior and neurophysiology in both naturalistic conditions and designed lab experiments. Unlike psychological theories, artificial neural networks (ANNs) actually perform visual tasks and generate testable predictions for arbitrary inputs. These advantages enable ANNs to engage the entire spectrum of the evidence. Failures of particular models drive progress in a vibrant ANN research program of human vision.

Bowers et al. discuss the limited connection between the psychological literature on human vision and recent work combining artificial neural networks (ANNs) and benchmark-based statistical evaluation. They are correct that the psychological literature has described behavioral signatures of human vision that ANNs should but do not currently explain. A model of human vision should ideally explain all available neural and behavioral data, including the unprecedentedly rich data from naturalistic benchmarks as well as data from experiments designed to address specific psychological hypotheses. None of the current models (ANNs, handcrafted computational models, and abstractly described psychological theories) meet this challenge.

Importantly, however, the failure of current ANNs to explain all available data does not amount to a refutation of neural network models in general. Falsifying the entire, highly expressive class of ANN models is impossible. ANNs are universal approximators of dynamical systems (Funahashi & Nakamura, 1993; Schäfer & Zimmermann, 2007) and hence can implement any potential computational mechanism. Future ANNs may contain different computational mechanisms that have not yet been explored. ANNs therefore are best understood not as a monolithic falsifiable *theory* but as a *computational language* in which particular falsifiable hypotheses can be expressed. Bowers et al.'s long list of cited studies presenting shortcomings of particular models neither demonstrates the failure of the ANN modeling framework in general nor a lack of openness of the field to falsifications of ANN models. Instead, their list of citations rather impressively illustrates the opposite: That the emerging ANN research program (referred to as "neuroconnectionism" in Doerig et al., 2022) is progressive in the sense of Lakatos: It generates a rich variety of falsifiable hypotheses (expressed in the language of ANNs) and advances through model comparison (Doerig et al., 2022). Each shortcoming drives improvement. For example, the discovery of texture bias in ANNs (Geirhos et al., 2019) has led to a variety of alternative training methods that make ANNs rely more strongly on larger-scale structure in images (e.g., Geirhos et al., 2019; Hermann, Chen, & Kornblith, 2020; Nuriel, Benaim, & Wolf, 2021). Similarly, the discovery of adversarial susceptibility of ANNs (Szegedy et al., 2013) has motivated much research on perceptual robustness (e.g., Cohen, Rosenfeld, & Kolter, 2019; Guo et al., 2022; Madry, Makelov, Schmidt, Tsipras, & Vladu, 2019).

Bowers et al. create a false dichotomy between benchmark studies (e.g., Cichy, Roig, & Oliva, 2019; Kriegeskorte et al., 2008; Nonaka, Majima, Aoki, & Kamitani, 2021; Schrimpf et al., 2018) and controlled psychological experiments. Both approaches test model-based predictions of empirical data. Traditional psychological experiments are designed to test verbally defined theories, minimizing confounders of the independent variables of theoretical interest. In contrast, the numerous experimental conditions included in natural image behavioral and neural benchmarks are high-dimensional, complex, and ecologically relevant. Controlled experiments pose specific questions. They promise to give us theoretically important bits of information but are biased by theoretical assumptions and risk missing the computational challenge of task performance under realistic conditions (Newell, 1973; Olshausen & Field, 2005). Observational studies and experiments with large numbers of natural images pose more general questions. They promise evaluation of many models with comprehensive data under more naturalistic conditions, but risk inconclusive results because they are not designed to adjudicate among alternative computational mechanisms (Rust & Movshon, 2005). Between these extremes lies a rich space of neural and behavioral empirical tests for models of vision. The community should seek models that can account for data across this spectrum, not just one end of it.

Despite their widely discussed shortcomings (e.g., Lindsay, 2021; Peters & Kriegeskorte, 2021; Serre, 2019), ANNs are sometimes referred to as the "current best" models of human vision. This characterization is justified on both a priori and empirical grounds. A priori, ANNs are superior to verbally defined cognitive theories in that they are image-computable, that is, they are fully computationally specified and take images as input. These properties enable ANNs to make quantitative predictions about a broad range of empirical phenomena, rendering ANNs more amenable to falsification. Being fully computationally specified enables them to make quantitative predictions of neural and behavioral responses (an advantage shared with other cognitive computational models). Taking images as inputs enables ANNs to make predictions about neural and behavioral responses to arbitrary visual stimuli. A model that explains only a particular psychological phenomenon is a priori inferior, ceteris paribus, to a model that predicts data across a wide range of conditions and dependent measures. The discrepancies between human vision and current ANNs are "bugs" of particular models, but the fact that we can discover these bugs is a feature of image-computable ANNs, fueling empirical progress. Since ANNs are image-computable, they enable severe tests of their predictions (superstimuli, adversarial examples, metamers; Bashivan, Kar, & DiCarlo, 2019; Dujmović, Malhotra, & Bowers, 2020; Feather, Durango, Gonzalez, & McDermott, 2019; Walker et al., 2019) and powerful model comparisons (controversial stimuli; Golan, Raju, & Kriegeskorte, 2020).

The empirical reason why ANNs can be called the "current best" models of human vision is that they offer unprecedented mechanistic explanations of the human capacity to make sense of complex, naturalistic inputs. Most basically, ANNs are currently the only models that can recognize objects, parse scenes, or identify faces at performance levels similar to human performance. Furthermore, they offer image-specific predictions of errors (e.g., Geirhos et al., 2021; Rajalingham et al., 2018) and reaction times (e.g., Spoerer, McClure, & Kriegeskorte, 2017). Their predictions are far from perfect but better than those of alternative models. Finally, the intermediate representations of ANNs currently best match the neural representations that underlie human visual capacities (e.g., Dwivedi, Bonner, Cichy, & Roig, 2021; Güçlü & van Gerven, 2015).

In sum, ANNs provide a language that enables us to express and test falsifiable computational models that have extraordinary power and can generalize to a broad range of empirical phenomena. Lakatos (1978) noted that all theories "are born refuted and die refuted" and stressed the importance of comparing competing theories in the light of the evidence. Our studies, then, should

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

compare many models and report both their failures and their relative successes. It is through creation and comparison of many models that our field will progress.

Financial support. This research received no specific funding from any funding agency or commercial or not-for-profit entity.

Competing interest. None.

References

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science (New York, N.Y.)*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Cichy, R. M., Roig, G., & Oliva, A. (2019). The Algonauts project. *Nature Machine Intelligence*, *1*(12), 613–613. https://doi.org/10.1038/s42256-019-0127-z

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning*. Proceedings of Machine Learning Research, Long Beach, CA, USA (Vol. 97, pp. 1310–1320). https://proceedings.mlr.press/v97/cohen19c.html

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., … Kietzmann, T. C. (2022). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*, 431–450. https://doi.org/10.1038/s41583-023-00705-w

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

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, *17*(8), e1009267. https://doi.org/10.1371/journal.pcbi.1009267

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'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada (Vol. 32, pp. 10078–10089). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/ac27b77292582bc293a51055bfc994ee-Paper.pdf

Funahashi, K. I., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, *6*(6), 801–806. https://doi.org/10.1016/S0893-6080(05)80125-X

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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 23885–23899). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf

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

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Guo, C., Lee, M., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & Dicarlo, J. (2022). Adversarially trained neural representations are already as robust as biological neural representations. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning*. Proceedings of Machine Learning Research, Baltimore, MD, USA (Vol. 162, pp. 8072–8081). PMLR. https://proceedings.mlr.press/v162/guo22d.html

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada (Vol. 33, pp. 19000–19015). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/db5f9f42a7157abe65bb145000b5871a-Paper.pdf

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. https://doi.org/10.1016/j.neuron.2008.10.043

Lakatos, I. (1978). Science and pseudoscience. *Philosophical Papers*, *1*, 1–7.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). Towards deep -learning models resistant to adversarial attacks. In *International conference on learning representations*, Vancouver, BC, Canada. https://openreview.net/forum?id=rJzIBfZAb

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), Visual information processing: *Proceedings of the 8th annual Carnegie symposium on cognition, held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972* (pp. 283–305). Academic Press.

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like?. *iScience*, *24*(9), 103013. https://doi.org/10.1016/j.isci.2021.103013

Nuriel, O., Benaim, S., & Wolf, L. (2021). Permuted AdaIN: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9482–9491). Online. https://openaccess.thecvf.com/content/CVPR2021/html/Nuriel_Permuted_AdaIN_Reducing_the_Bias_Towards_ Global_Statistics_in_Image_CVPR_2021_paper.html

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1?. *Neural Computation*, *17*(8), 1665–1699. https://doi.org/10.1162/0899766054026639

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*(9), 1127–1144. https://doi.org/10.1038/s41562-021-01194-6

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. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*(12), 1647–1650. https://doi.org/10.1038/nn1606

Schäfer, A. M., & Zimmermann, H. G. (2007). Recurrent neural networks are universal approximators. *International Journal of Neural Systems*, *17*(4), 253–263. https://doi.org/10.1142/S0129065707001111

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., … DiCarlo, J. J. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-like?. *BioRxiv*, 407007. https://doi.org/10.1101/407007

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

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551. https://doi.org/10.3389/fpsyg.2017.01551

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv. arXiv:1312.6199*. https://doi.org/10.48550/arXiv.1312.6199

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., … Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

There is a fundamental, unbridgeable gap between DNNs and the visual cortex

Moshe Gur

Department of Biomedical Engineering, Technion, Haifa, Israel
mogi@bm.technion.ac.il

doi:10.1017/S0140525X23001590, e393

Abstract

Deep neural networks (DNNs) are not just inadequate models of the visual system but are so different in their structure and functionality that they are not even on the same playing field. DNN units have almost nothing in common with neurons, and, unlike visual neurons, they are often fully connected. At best, DNNs can label inputs, while our object perception is both holistic and detail preserving. A feat that no computational system can achieve.

The authors make a valuable contribution in pointing out that deep neural networks (DNNs) are not good models of the visual system

since they rely on predictions and fail to account for results from many psychophysical experiments. However, the authors' implicit acceptance that DNN basic structure and operational principles are a fair simulation of the visual system resulted in ignoring that DNNs are not just inadequate to represent biological vision but that they are not even on the same playing field.

The discovery of feature-selective cells in primates V1 and subsequent findings of face-selective cells in the monkey inferotemporal (IT) cortex have led to the dominant physiology-based view that object representation and recognition is encoded by sparse populations of expert cell ensembles (Freiwald & Tsao, 2010). This theory posits that, first, the image is analyzed into its basic elements, such as line segments, by V1 feature-selective cells. Then, after hierarchical convergence and integration of simple elements, expert cell ensembles, by their collective responses, represent objects uniquely.

Those physiological findings and models have inspired various computational models (Marr, 2010), leading to current DNNs. Such models, though incorporating some neuronal-like characteristics, were not intended to emulate brain mechanisms; much of the perceived similarity between biological and artificial networks has more to do with terminology (e.g., cells, layers, learning, features) than with exact replication. The fundamental differences between biological and artificial object recognition mechanisms are manifested in functional anatomy, and, most importantly, in the inherent impossibility of DNNs, or any other computational mechanism, including expert ensembles, to replicate our perceptual abilities.

Functional anatomy

Profound differences between biological networks and DNNs can be found at all organizational levels. Single neurons with their complex electro-chemical interactions are unlike the schematic DNN "neurons." As noted by Ullman (2019), almost everything known about biological neurons was left out of their DNN counterparts. Connectivity between single elements presents another striking divergence between DNNs and the visual cortex. Typically, DNNs contain many layers with units that are connected to all other units, although there is no such connectivity in the visual cortex. Lateral connections within a cortical area are sparse; in V1 the intrinsic connections are only between same orientation columns (Stettler, Aniruddha, Bennett, & Gilbert, 2002). Feed-forward connections in the ventral pathway, V1 → V2 → V4 → IT, are not one-to-one but many-to-one leading to increase in receptive field size from minutes of arc in V1 to many degrees in the IT cortex (Rolls, 2012). Most feedback from higher to lower areas is less dense than the feed-forward one and does not target single cells (Rolls, 2012). Also, DNNs are usually composed of more than 20 layers including highly specialized ones

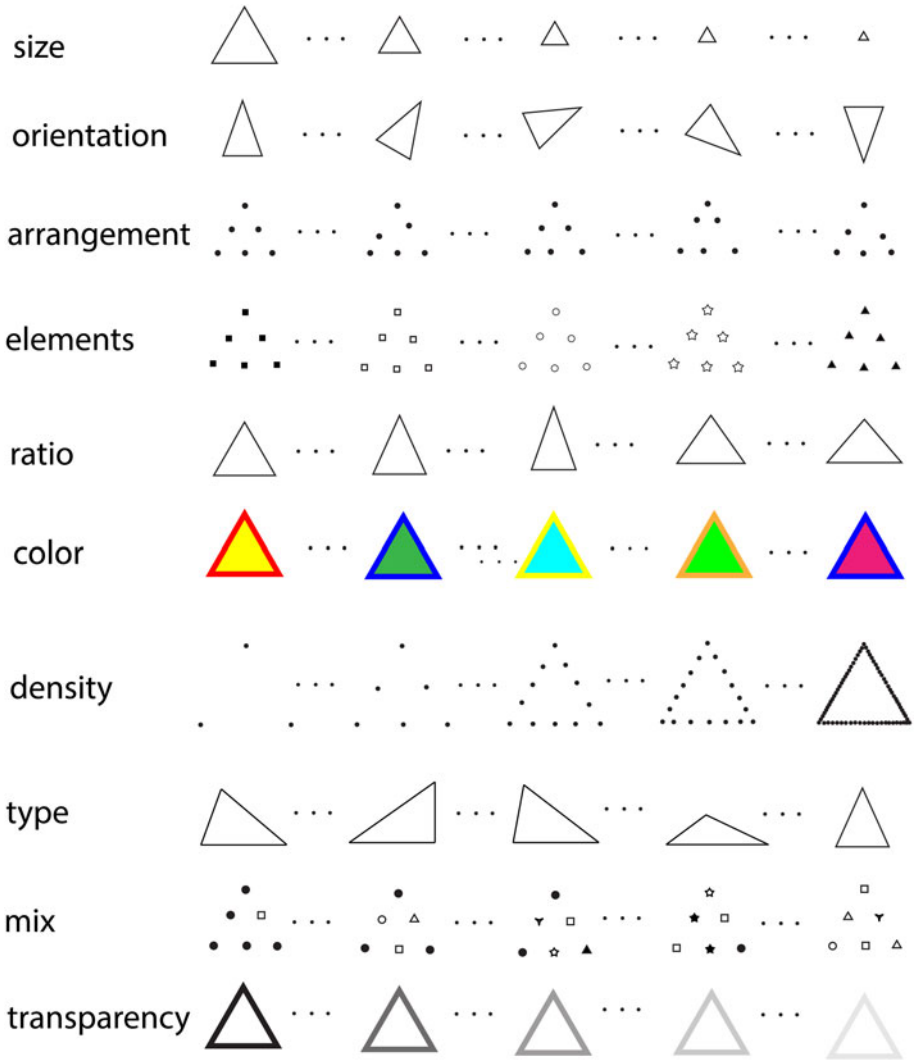


Figure 1 (Gur). Triangle is recognized as such despite potentially appearing in any one of an astronomical number of variations.

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

whereas there are only four areas in the visual hierarchy which are remarkably similar in their basic anatomy. While such profound structural differences are in themselves sufficient to deny that DNNs "are the best models of biological vision," even a larger gap between DNNs and biological reality stems from ultimate differences in functionality. Note that since both expert ensembles and DNN models share essential characteristics, such as hierarchical feature extraction and integration, and since the latter differ structurally from the visual cortex, showing that the biologically based model is not a good model of the visual cortex, rules out DNNs as well.

Invariance

We recognize an object even though it can vary greatly in appearance. In Figure 1 a triangle with many properties (e.g., size) and many variations within property is displayed. Combining properties and variations leads to an astronomical number of variations, yet all objects are recognized as triangles, are differentiated from each other, and from any of the similarly astronomical possible appearances of, say, a square. It is impossible for a small number of "expert" cells to generate a pattern that is unique to every single triangle yet differs from all squares-generated patterns. The same arguments apply, obviously, to all other object categories. Are we to assume then that there are ensembles dedicated to each of the many thousands object categories out there?

Space and time characteristics of object perception

A collection (~0.6° in total width) of small dots displayed for ~50 ms is easily perceived as a turtle (Fig. 2; Gur, 2021). This result is quite informative; the display size means that the dots must be represented by V1 cells which are the only ones with small enough receptive fields. The short exposure means that correctly judging the position and relationship between dots, which is essential for a holistic object perception, cannot result from V1 → V2 → V4 → IT hierarchical convergence since there is simply not enough time (Schmolesky et al., 1998). This example is consistent with many studies showing an accurate perception of flashed objects (cf. Crouzet, Kirchner, & Thorpe, 2010; Greene & Visani, 2015; Gur, 2018, 2021; Keysers, Xiao, Foldiak, & Perrett, 2001). Thus, unlike the predictions of the expert ensemble theory, our perception is almost instantaneous, parallel, and detail preserving.

Detailed and holistic

These two characteristics are contradictory to any integrative/computational system where the whole is derived by integrating over its parts. However, this is how we see the world. It is the visual system ability to perceive space simultaneously and in parallel that leads to the holistic yet detailed capacity. The world is perceived almost in a flash-like manner; all the elements' position and features (size, shape, orientation, etc.) are available for an immediate decision – a building or a face? Such a discrimination between objects with such a retention of details would not have been possible if elements were integrated and serially sent across many synapses to areas downstream from V1.

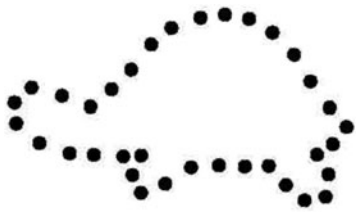


Figure 2 (Gur). Animal is recognized even when its contour is represented by dots, and is displayed for ~50 ms.

Finally, any computational system, including those discussed here, can, at best, map output to input. To recognize an object, the system generates a label, "object #2319764," which is devoid of all the object's constituent elements. This result is quite different from our perceptual experience where an object is perceived together with its minute details. This discrepancy between perceptual reality and even the best possible performance of a computational system clearly means that such a system cannot be a model of the visual system.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*, 10–17.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face processing system. *Science (New York, N.Y.)*, *330*, 845–851.

Greene, E., & Visani, A. (2015). Recognition of letters displayed as briefly flashed dot patterns. *Attention Perception & Psychophysics*, *77*, 1955–1969.

Gur, M. (2018). Very small faces are easily discriminated under long and short exposure times. *Journal of Neurophysiology*, *119*, 1599–1607. doi:10.1152/jn.00622.2017

Gur, M. (2021). Psychophysical evidence and perceptual observations show that object recognition is not hierarchical but is a parallel, simultaneous, egalitarian, non-computational system. *BioRxiv*, 1–25. doi:10.1101/2021.06.10.447325

Keysers, C., Xiao, D.-K., Foldiak, P., & Perrett, D. I. (2001). The speed of light. *Journal of Cognitive Neuroscience*, *13*, 90–101.

Marr, D. (2010). *Vision*. MIT Press.

Rolls, E. T. (2012). Invariant visual object and face recognition: Neural and computational bases, and a model, VisNet. *Frontiers in Computational Neuroscience*, *6*, 1–70.

Schmolesky, M. T., Youngchang, W., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., & Leventhal, A. G. (1998). Signal timing in the macaque visual system. *Journal of Neurophysiology*, *79*, 3272–3278.

Stettler, D. D., Aniruddha, D., Bennett, J., & Gilbert, C. D. (2002). Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, *36*, 739–750.

Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science (New York, N.Y.)*, *363*, 692–693.

For human-like models, train on human-like tasks

Katherine Hermann[a], Aran Nayebi[b], Sjoerd van Steenkiste[c] and Matt Jones[c,d]

[a]Google DeepMind, Mountain View, CA, USA; [b]McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA; [c]Google Research, Mountain View, CA, USA and [d]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO, USA

hermannk@google.com
aran.nayebi@gmail.com
sjoerdvansteenkiste@gmail.com
mcj@colorado.edu
https://anayebi.github.io/
https://www.sjoerdvansteenkiste.com/
http://matt.colorado.edu

doi:10.1017/S0140525X23001516, e394

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

Abstract

Bowers et al. express skepticism about deep neural networks (DNNs) as models of human vision due to DNNs' failures to account for results from psychological research. We argue that to fairly assess DNNs, we must first train them on more human-like tasks which we hypothesize will induce more human-like behaviors and representations.

We agree with Bowers et al. that accounting for results from behavioral experiments should serve as a North Star as we develop models of human vision. But what is a promising path to finding models that perform well on experimental benchmarks? In this commentary, we focus on the role of the task(s) on which models are trained. Zhang, Bengio, Hardt, Recht, and Vinyals (2017) have shown that modern deep neural networks (DNNs) are more than expressive enough to overfit to any classification task on which they are trained. In particular, the authors show that DNNs can learn to classify ImageNet images (Deng et al., 2009) with arbitrarily shuffled labels, demonstrating maximal flexibility with respect to this training set. To introduce a metaphor, our models are like sponges, capable of absorbing whatever information we teach them through the training tasks we present. Thus, when we ask about a model's behavior, we should ask, first, what it was trained to do. Although Bowers et al. take failures of ImageNet-trained models to behave in human-like ways as support for abandoning DNN architectures, we argue that we should instead consider alternative training tasks for DNNs.

Recent work has shown that pushing DNNs to perform well on ImageNet may not, in general, push them to be more human-like. Very high ImageNet performance becomes inversely related to primate neural predictivity (Schrimpf et al., 2018; Schrimpf, 2022), and there is a tradeoff between perceptual scores from human judgments (Kumar, Houlsby, Kalchbrenner, & Cubuk, 2022) and ImageNet performance, and between shape bias and ImageNet performance, when shape bias is modulated by data augmentation (Hermann, Chen, & Kornblith, 2020).

Certainly, humans can categorize the objects they see, but categorization is only a small part of how we process the visual world. Mostly, we use our visual systems to *interact* with the objects around us, in a closed loop comprising perception, inference, decision making, and action. There are several reasons to believe that training models on similarly embodied and active learning tasks may bring their behavior and representations closer to humans'. First, physically interacting with objects requires detailed perception of their global spatial properties (shape, position, motor affordances, etc.). Arguably, several of the most famous divergences between models and people stem from models' failures to weigh exactly this kind of information. For example, unlike people (Kucker et al., 2019; Landau, Smith, & Jones, 1988), many standard DNNs seem to rely on texture information more than shape (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos et al., 2019; Hermann et al., 2020). While, empirically, texture seems to be sufficient for good performance on ImageNet, it is unlikely to suffice for embodied navigation or manipulation tasks. In determining how to position oneself to sit in a chair, the shape and position of the chair are far more important than its color or upholstery texture. Similarly, adversarial examples (Nguyen, Yosinski, & Clune, 2015; Szegedy et al., 2013), another often-cited separator of humans and DNNs, arguably arise from models' over-reliance on local pixel patterns at the expense of the global configural information required for embodied interaction. Overall, we hypothesize that existing DNN architectures, if trained to navigate the world and interact with objects in the way that humans do, would be more likely to display human-like visual behavior and representations than they do under current training methods.

Another implication of the Zhang et al. (2017) work is that modern networks are sufficiently large that training them on a 1,000-way classification task on a million images is insufficient to exhaust their capacity, leaving important degrees of freedom governing their generalization performance underconstrained, which allows for deviant phenomena such as adversarial examples of the kind and severity currently observed. As another example of flexibility in how DNNs can learn a classification task, models often learn spurious/shortcut features (Arjovsky, Bottou, Gulrajani, & Lopez-Paz, 2019; Geirhos et al., 2020; McCoy, Pavlick, & Linzen, 2020), for example, using image backgrounds rather than foreground objects (Beery, Van Horn, & Perona, 2018; Xiao, Engstrom, Ilyas, & Madry, 2021), or single diagnostic pixels rather than other image content (Malhotra, Evans, & Bowers, 2020). This brings us to a second argument in favor of embodied training tasks. A dataset of similar size to ImageNet but with a richer, more ecological output space – for example, choosing a physical action and its control parameters, or predicting subsequent frames – would contain a vastly larger amount of information, perhaps more fully constraining the model's behavior.

Existing work validates the impact of training tasks on model behavior and representations. Even when restricted to training on ImageNet images, the training objective and/or data augmentation can affect how well models match human similarity judgments of images (Muttenthaler, Dippel, Linhardt, Vandermeulen, & Kornblith, 2023), categorization patterns (Geirhos et al., 2021), performance on real-time and life-long learning benchmarks (Zhuang et al., 2022), and feature preferences (Hermann et al., 2020), and also how well they predict primate physiology (Zhuang et al., 2021) and human fMRI (Konkle & Alvarez, 2022) data. Still, it is possible to enrich DNN training tasks much further, even for object categorization (Sun, Shrivastava, Singh, & Gupta, 2017).

We have discussed the promise of training embodied, interactive agents in rich, ethologically relevant environments. What efforts have already been made in this direction, and what might they look like in the future? Past work situating a vision system within a simulated agent navigating and interacting with its environment gives promising initial indications that human-like visual behaviors can emerge in this setting (Haber, Mrowca, Wang, Fei-Fei, & Yamins, 2018; Hill et al., 2020; Nayebi et al., 2021; Weihs et al., 2021). The continued development of new, more naturalistic training environments (Gan et al., 2021; Greff et al., 2022; Puig et al., 2018; Savva et al., 2019; Xiang et al., 2020) should support pushing this research program still further toward human-like learning. In addition, state-of-the-art large language models provide a new means of communicating richer tasks to models (Chen et al., 2023), and a new reservoir of human-like knowledge for models to draw on (Brohan et al., 2023). We predict further work in these directions will address shortcomings Bowers et al. identify and yield improved DNN accounts of human vision.

Acknowledgments. We thank Mike Mozer and Robert Geirhos for interesting discussions and helpful feedback.

Financial support. Aran Nayebi is supported by a K. Lisa Yang Integrative and Computational Neuroscience (ICoN) Postdoctoral Fellowship. Matt Jones is supported in part by NSF Grant 2020906.

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

Competing interest. None.

References

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

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.

Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 456–473).

Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., … Fu, C. K. (2023). Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on robot learning* (pp. 287–318). PMLR.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., … Soricut, R. (2023). Pali: A jointly-scaled multilingual language-image model. *International conference on learning representations (ICLR)*.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., … Yamins, D. L. K. (2021). ThreeDWorld: A platform for interactive multi-modal physical simulation. *Advances in Neural Information Processing Systems (NeurIPS)*.

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., 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. *Advances in Neural Information Processing Systems (NeurIPS)*, *34*, 23885–23899.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International conference on learning representations (ICLR)*.

Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., … Tagliasacchi, A. (2022). Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3749–3761).

Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F., & Yamins, D. L. (2018). Learning to play with intrinsically-motivated, self-aware agents. *Advances in Neural Information Processing Systems (NeurIPS)*, *31*.

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

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). Environmental drivers of systematicity and generalization in a situated agent. *International conference on learning representations (ICLR)*.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491.

Kucker, S. C., Samuelson, L. K., Perry, L. K., Yoshida, H., Colunga, E., Lorenz, M. G., & Smith, L. B. (2019). Reproducibility and a unifying explanation: Lessons from the shape bias. *Infant Behavior and Development*, *54*, 156–165.

Kumar, M., Houlsby, N., Kalchbrenner, N., & Cubuk, E. D. (2022). Do better ImageNet classifiers assess perceptual similarity better? *Transactions of Machine Learning Research*.

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

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.

McCoy, R. T., Pavlick, E., & Linzen, T. (2020). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th annual meeting of the association for computational linguistics, ACL 2019* (pp. 3428–3448). Association for Computational Linguistics (ACL). https://aclanthology.org/P19-1334/

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023). Human alignment of neural network representations. *International conference on learning representations (ICLR)*.

Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, A. M., & Yamins, D. L. (2021). Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *BioRxiv*, 2021-06.

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).

Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., & Torralba, A. (2018). VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8494–8502).

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., … Batra, D. (2019). Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9339–9347).

Schrimpf, M. (2022). Advancing system models of brain processing via integrative benchmarking. Doctoral dissertation, Massachusetts Institute of Technology.

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

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (pp. 843–852).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Weihs, L., Kembhavi, A., Ehsani, K., Pratt, S. M., Han, W., Herrasti, A., … Farhadi, A. (2021). Learning generalizable visual representations via interactive gameplay. *International conference on learning representations (ICLR)*.

Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., … Su, H. (2020). Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11097–11107).

Xiao, K., Engstrom, L., Ilyas, A., & Madry, A. (2021). Noise or signal: The role of image backgrounds in object recognition. *International conference on learning representations (ICLR)*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *International conference on learning representations (ICLR)*.

Zhuang, C., Xiang, V., Bai, Y., Jia, X., Turk-Browne, N., Norman, K., … Yamins, D. L. (2022). How well do unsupervised learning algorithms model human real-time and life-long learning? In Thirty-sixth conference on neural information processing systems datasets and benchmarks track.

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 of the United States of America*, *118*(3), e2014196118.

Beyond the limitations of any imaginable mechanism: Large language models and psycholinguistics

Conor Houghton[a], Nina Kazanina[b,c] and Priyanka Sukumaran[a,b]

[a]Department of Computer Science, University of Bristol, Bristol, UK; [b]School of Psychological Sciences, University of Bristol, Bristol, UK and [c]International Laboratory of Social Neurobiology, Institute for Cognitive Neuroscience, National Research University, Higher School of Economics, HSE University, Moscow, Russia

conor.houghton@bristol.ac.uk
nina.kazanina@bristol.ac.uk
p.sukumaran@bristol.ac.uk
conorhoughton.github.io

doi:10.1017/S0140525X23001693, e395

Abstract

Large language models (LLMs) are not detailed models of human linguistic processing. They are, however, extremely successful at their primary task: Providing a model for language. For this reason LLMs are important in psycholinguistics: They are useful as a practical tool, as an illustrative comparative, and philosophically, as a basis for recasting the relationship between language and thought.

Neural-network models of language are optimized to solve practical problems such as machine translation. Currently, when these large language models (LLMs) are interpreted as models of human linguistic processing they have similar shortcomings to

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

those that deep neural networks have as models of human vision. Two examples can illustrate this. First, LLMs do not faithfully replicate human behaviour on language tasks (Kuncoro, Dyer, Hale, & Blunsom, 2018; Linzen & Leonard, 2018; Marvin & Linzen, 2018; Mitchell, Kazanina, Houghton, & Bowers, 2019). For example, an LLM trained on a word-prediction task shows similar error rates to humans overall on long-range subject–verb number agreement but errs in different circumstances: Unlike humans, it makes more mistakes when sentences have relative clauses (Linzen & Leonard, 2018), indicating differences in how grammatical structure is represented. Second, the LLMs with better performance on language tasks do not necessarily have more in common with human linguistic processing or more obvious similarities to the brain. For example, transformers learn efficiently on vast corpora and avoid human-like memory constraints but are currently more successful as language models than recurrent neural networks such as the long- and short-term memory LLMs (Brown et al., 2020; Devlin, Chang, Lee, & Toutanova, 2018), which employ sequential processing, as humans do, and can be more easily compared to the brain.

Furthermore, the target article suggests that, more broadly, the brain and neural networks are unlikely to resemble each other because evolution differs in trajectory and outcome from the optimization used to train a neural network. Generally, there is an unanswered question about which aspects of learning in LLMs are to be compared to the evolution of our linguistic ability and which to language learning in infants but in either case, the comparison seems weak. LLMs are typically trained using a next-word prediction task; it is unlikely our linguistic ability evolved to optimize this and next-word prediction can only partly describe language learning: For example, infants generalize word meanings based on shape (Landau, Smith, & Jones, 1988) while LLMs lack any broad conceptual encounter with the world language describes.

In fact, it would be peculiar to suggest that LLMs are models of the neural dynamics that support linguistic processing in humans; we simply know too little about those dynamics. The challenge presented by language is different to that presented by vision: Language lacks animal models and debate in psycholinguistics is occupied with broad issues of mechanisms and principles, whereas visual neuroscience often has more detailed concerns. We believe that LLMs have a valuable role in psycholinguistics and this does not depend on any precise mapping from machine to human. Here we describe three uses of LLMs: (1) the *practical*, as a tool in experimentation; (2) the *comparative*, as an alternate example of linguistic processing; and (3) the *philosophical*, recasting the relationship between language and thought.

(1) An LLM models language and this is often of *practical* quantitative utility in experiment. One straight-forward example is the evaluation of *surprisal*: How well a word is predicted by what has preceded it. It has been established that reaction times (Fischler & Bloom, 1979; Kleiman, 1980), gaze duration (Rayner & Well, 1996), and EEG responses (Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Frank, Otten, Galli, & Vigliocco, 2015) are modulated by surprisal, giving an insight into prediction in neural processing. In the past, surprisal was evaluated using *n*-grams, but *n*-grams become impossible to estimate as *n* grows and as such they cannot quantify long-range dependencies. LLMs are typically trained on a task akin to quantifying surprisal and are superior to *n*-grams in estimating word probabilities. Differences between LLM-derived estimates and neural perception of surprisal may quantify which linguistic structures, perhaps poorly represented in the statistical evidence, the brain privileges during processing.

(2) LLMs are also useful as a point of *comparison*. LLMs combine different computational strategies, mixing representations of word properties with a computational engine based on memory or attention. Despite the clear differences between LLMs and the brain, it is instructive to compare the performance of different LLMs on language tasks to our own language ability. For example, although LLMs are capable of long-range number and gender agreement (Bernardy & Lappin, 2017; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Linzen, Dupoux, & Goldberg, 2016; Sukumaran, Houghton, & Kazanina, 2022), they are not successful in implementing another long-range rule: Principle C (Mitchell et al., 2019), a near-universal property of languages which depends in its most straight-forward description on hierarchical parsing. Thus, LLMs allow us to recognize those aspects of language which require special consideration while revealing others to be within easy reach of statistical learning.

(3) In the past, *philosophical* significance was granted to language as evidence of thought or personhood. Turing (1950), for example, proposes conversation as a proxy for thought and Chomsky (1966) describes Descartes as attributing the possession of mind to other humans because the human capacity for innovation and for the creative use of language is "beyond the limitations of any imaginable mechanism." It is significant that machines are now capable of imitating the use of language. While machine-generated text still has attributes of awkwardness and repetition that make it recognizable on careful reading, it would seem foolhardy to predict these final quirks are unresolvable or are characteristic of the division between human and machine. Nonetheless, most of us appear to feel intuitively that LLMs enact an imitation rather than a recreation of our linguistic ability: LLMs seem empty things whose pantomime of language is not underpinned by thought, understanding, or creativity. Indeed, even if an LLM were capable of imitating us perfectly, we would still distinguish between a loved one and their simulation.

This is a challenge to our understanding of the relationship between language and thought: Either we must claim that, despite recent progress, machine-generated language will remain unlike human language in vital respects, or we must defy our intuition and consider machines to be as capable of thought as we are, or we must codify our intuition to specify why a machine able to produce language should, nonetheless, be considered lacking in thought.

Acknowledgments. We are grateful to the many colleagues who read and commented on this text.

Financial support. P. S. received support from the Wellcome Trust [108899/B/15/Z], C. H. from the Leverhulme Trust [RF-2021-533], and N. K. from the International Laboratory for Social Neurobiology of the Institute for Cognitive Neuroscience HSE, RF Government grant [075-15-2022-1037].

Competing interest. None.

References

Bernardy, J.-P., & Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, *15*(2), 1–15.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

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

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Cambridge University Press.
Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (pp. 4171–4186). https://arxiv.org/abs/1810.04805
Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 1–20.
Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.
Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1195–2205). https://arxiv.org/pdf/1803.11138.pdf
Kleiman, G. M. (1980). Sentence frame contexts and lexical decisions: Sentence-acceptability and word- relatedness effects. *Memory & Cognition*, *8*(4), 336–344.
Kuncoro, A., Dyer, C., Hale, J., & Blunsom, P. (2018). The perils of natural behaviour tests for unnatural models: The case of number agreement. Poster presented at learning language in humans and in machines, Paris, France, July, 5(6). Organizers: Susan Goldin-Meadow, Afra Alishahi, Phil Blunsom, Cynthia Fisher, Chen Yu & Michael Frank.
Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.
Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.
Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In C. Kalish, M. A. Rau, X. (J.) Zhu, & T. T. Rogers (Eds.), *Proceedings of CogSci 2018* (pp. 692–697).
Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202).
Mitchell, J., Kazanina, N., Houghton, C., & Bowers, J. (2019). Do LSTMs know about Principle C?. In H. Nienborg, R. Poldrack, & T. Naselaris (Eds.), *Conference on Cognitive Computational Neuroscience*, Berlin. https://doi.org/10.32470/CCN.2019.1241-0
Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509.
Sukumaran, P., Houghton, C., & Kazanina, N. (2022). Do LSTMs see gender? Probing the ability of LSTMs to learn abstract syntactic rules. In J. Bastings, Y. Belinkov, Y. Elazar, D. Hupkes, N. Saphr, & S. Wiegreffe (Eds.), *Poster at BlackboxNLP 2022*. https://arxiv.org/abs/2211.00153
Turing, A. M. (1950). Computing machinery and intelligence. *Mind; A Quarterly Review of Psychology and Philosophy*, *49*, 433–460.

For deep networks, the whole equals the sum of the parts

Philip J. Kellman[a], Nicholas Baker[b], Patrick Garrigan[c], Austin Phillips[d] and Hongjing Lu[e]

[a]Department of Psychology and David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [b]Department of Psychology, Loyola University of Chicago, Chicago, IL, USA; [c]Department of Psychology, St. Joseph's University, Philadelphia, PA, USA; [d]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA and [e]Department of Psychology and Department of Statistics, University of California, Los Angeles, Los Angeles, CA, USA
kellman@cognet.ucla.edu; https://kellmanlab.psych.ucla.edu/
nbaker1@ucla.edu; https://www.luc.edu/psychology/people/staff/facultyandstaff/nicholasbaker/
patrick.garrigan@sju.edu; https://sjupsych.org/faculty_pg.php
asphillips@ucla.edu; https://kellmanlab.psych.ucla.edu/
hongjing@ucla.edu; https://cvl.psych.ucla.edu/

doi:10.1017/S0140525X23001541, e396

Abstract

Deep convolutional networks exceed humans in sensitivity to local image properties, but unlike biological vision systems, do not discover and encode abstract relations that capture important properties of objects and events in the world. Coupling network architectures with additional machinery for encoding abstract relations will make deep networks better models of human abilities and more versatile and capable artificial devices.

Bowers et al. raise questions about the validity of methods and types of evidence used to compare deep networks to human vision. Their discussion also draws attention to serious limitations of convolutional neural networks for understanding vision. Here we focus on two ideas. First, compelling evidence is emerging that deep networks do not capture pervasive and powerful aspects of visual perceptual capabilities in humans. These limitations appear to be fundamental and relate to the lack of mechanisms for extracting and encoding *abstract relations*. Second, the problem of mimicry, both in comparing network and human responses in behavioral tasks and in comparing model unit activations to brain data, highlights a general difficulty in using potentially superficial similarities across systems to draw deep conclusions. We conclude by suggesting that understanding the mechanisms of visual perception will likely require synergies between network processing and processes that accomplish symbolic encoding of abstract relations.

Abstract relations in perception

Human perception derives abstract, symbolic representations from relational information in sensory input (e.g., Baker & Kellman, 2018), enabling visual representations to be widely useful in thinking and learning (Kellman & Massey, 2013). Processes like those found in deep convolutional neural networks (DCNNs) may be an important part of human vision, but their anchor in concrete, pixel-level properties makes them unlikely to be sufficient. DCNNs differ from human perceivers profoundly, for example, in their access to shape information (Baker, Lu, Erlikhman, & Kellman, 2018, 2020; Geirhos et al., 2019; Malhotra, Dujmović, & Bowers, 2022). Whereas shape is the pre-eminent driver of human object recognition, when shape and texture conflict, networks classify by texture. Humans readily see shape in glass figurines, but networks consistently misclassify these (e.g., labeling a robin as a shower cap, a fox as a chain, and a polar bear as a can opener). Silhouettes do better, producing around 40% accuracy (Baker et al., 2018; Kubilius, Bracci, & de Beeck, 2016), but rearranging their parts, which severely impairs human classification, has strikingly little effect on network responses. Conversely, for correctly classified silhouettes, adding small serrations along the boundary reduces network classifications to chance or below, while human perceivers are unaffected. These and other results indicate that networks extract local shape features but have little or no access to global shape (Baker et al., 2018, 2020).

Recent research suggests that these findings regarding global shape reflect broader limitations in DCNNs' abilities to capture abstract relations from visual inputs. Baker, Garrigan, Phillips, and Kellman (2023) attempted to train DCNNs to capture several perceptual relations that human perceivers detect readily and generalize from even a small number of examples. These included the same/different relation (Puebla & Bowers, 2021a, 2021b), judging

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

if a probe was inside or outside of a closed contour, and comparing the number of sides of two polygons. Using restricted and unrestricted transfer learning with networks previously trained for object classification, we found that networks could come to exceed chance performance on training sets. Subsequent testing with novel displays, however, showed that the relations *per se* were not learned at all. Although human perceivers rapidly acquired and accurately applied these relations to new displays, networks showed chance performance. The limitation of deep networks in representing and generalizing abstract relations appears to be fundamental and general (see also Malhotra, Dujmović, Hummel, & Bowers, 2021).

Methodological issues

The methodological issues raised by Bowers et al. are well-placed. Similarities between model responses and human judgments, and between model activations and brain activations, invite us to think that human processing may resemble deep networks. Yet claims based on both kinds of similarities may be tenuous. In our research, we have often seen deep networks produce somewhat better than chance responding on tests of relational processing, only to find that they were using some obscure, nonrelational property that correlated with the relevant relation.

Parallel concerns apply to similarities between activation patterns in brains and DCNN layers. Although intriguing, it is important to remember that we typically do not know what activations in *either* layers of a neural network or in the visual brain are signaling. Common representations in two systems might produce similar activation patterns or respectable correlations in representational similarity analyses (RSAs), but it is a case of affirming the consequent when we assume that high representational similarity scores imply common representations (Saxe, McClelland, & Ganguli, 2019). These issues may shed light on puzzling results. For example, Fan, Yamins, and Turk-Browne (2018) interpreted RSA results as suggesting that deep learning systems trained on photographs capture abstract representations such as humans use to see objects from line drawings. RSA was used to correlate similarity matrices obtained for photos and for line drawings; prior work used RSA to argue for quantitative similarities between advanced layers of the model and primate IT. In contrast, we tested classification of outline drawings by deep networks (VGG-19 and AlexNet) trained on ImageNet for object classification and found no evidence of successful classification based on outlines (Baker et al., 2018). For 78% of objects, networks would have done better by choosing an ImageNet category at random, and neither network produced a single correct first-choice classification. Do networks capture an abstract outline representation of objects, as suggested by RSA, yet fail to use it to classify inputs composed solely of outlines? As Bowers et al. suggest, the answer may lie in confounding in the stimulus properties that drive representational similarity (cf. Saxe et al., 2019).

Conclusion

A wealth of evidence suggests that biological vision systems extract and represent abstract relations. DCNNs far exceed humans in sensitivity to local image properties, but for humans, local sensory activations are transient, rapidly discarded, and used to discover and encode relations that capture important properties of objects and events in the world. Peering beyond observed similarities that Bowers et al. suggest may be superficial, these differences between networks and brains may be deep and fundamental. More work is needed to discern the sources of the differences. Combining network architectures with additional machinery for encoding abstract relations might make deep networks better models of human abilities and more versatile and capable artificial devices.

Financial support. This work was supported by funding from the National Institutes of Health (P.. K., R01CA236791) and the National Science Foundation (H. L., BCS-2142269).

Competing interest. None.

References

Baker, N., Garrigan, P., Phillips, A., & Kellman, P. J. (2023). Configural relations in humans and deep convolutional neural networks. *Frontiers in Artificial Intelligence*, *5*, 961595. doi:10.3389/frai.2022.961595

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*(9), 1295.

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.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research*, *172*, 46–61.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*(8), 2670–2698.

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

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 58, pp. 117–165). Elsevier.

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

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

Malhotra, G., Dujmović, M., Hummel, J., & Bowers, J. 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

Puebla, G., & Bowers, J. (2021a). Can deep convolutional neural networks support relational reasoning in the same-different task? *arXiv preprint*, https://doi.org/10.1101/2021.09.03.458919

Puebla, G., & Bowers, J. (2021b). Can deep convolutional neural networks learn same-different relations?. bioRxiv, 2021-04.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Modelling human vision needs to account for subjective experience

Marcin Koculak[a,b] and Michał Wierzchoń[a,b]

[a]Centre for Brain Research, Jagiellonian University, Krakow, Poland and [b]Consciousness Lab, Institute of Psychology, Jagiellonian University, Krakow, Poland

koculak.marcin@gmail.com
michal.wierzchon@uj.edu.pl
https://c-lab.pl

doi:10.1017/S0140525X2300170X, e397

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

Abstract

Vision is inseparably connected to perceptual awareness which can be seen as the culmination of sensory processing. Studies on conscious vision reveal that object recognition is just one of the means through which our representation of the world is built. We propose an operationalization of subjective experience in the context of deep neural networks (DNNs) that could encourage a more thorough comparison of human and artificial vision.

The target article comprehensibly deconstructs common misconceptions, such as models of human vision that can be reduced to mechanisms of object recognition or that useful analogies between neuronal and artificial architectures can be drawn solely from accuracy scores and their correlations with brain activity. We fully agree that such oversimplifications need to be avoided if deep neural networks (DNNs) are to be considered accurate models of vision. Troubles stemming from similar oversimplifications are well-known in consciousness research. One of the main obstacles for the field is the separation of mechanisms that process visual information from those that transform it into the conscious activity of seeing. Here, we offer a high-level outlook on the human vision from this perspective. We believe it could serve as a guiding principle for building more ecologically valid artificial models. It would also lead to better testing criteria for assessing the similarities and differences between humans and DNNs that go beyond object recognition.

When presented with an object, it appears that we first see it in all of its details and only then recognize it. However, experimental evidence suggests that, under carefully controlled conditions, individuals can correctly categorize objects while denying seeing them (Lamme, 2020). The discrepancy between objective performance (i.e., correct categorization) and subjective experience of seeing convincingly illustrates the presence of unconscious processing of perceptual information (Mudrik & Deouell, 2022). It also highlights that categorization may refer to different neural processes depending on the type of object. Identification of faces is a common example of fast automatic processing of a complex set of features that allows us to easily recognize each other. It also demonstrates problems with taking brain activity as an indicator of successful perception. The fusiform gyrus is selectively activated when participants are presented with images of faces (Fahrenfort et al., 2012; Haxby, Hoffman, & Gobbini, 2000). However, this activation can be found even if the participant reports no perception (Axelrod, Bar, & Rees, 2015). Similar specific neural activations can be observed in response to other complex stimuli (e.g., one's name) during sleep (Andrillon & Kouider, 2020). Therefore, while behavioural responses and brain activity can provide insights into the extent of processing evoked by certain stimuli, they do not equate to conscious vision.

Feature extraction and object categorization are not the only visual processes that can occur without consciousness. There is evidence of interactions between already differentiated objects that alter each other neural responses when placed closely in the visual field (Lamme, 2020). This includes illusions like the Kanizsa triangle, which requires the integration of multiple objects (Wang, Weng, & He, 2012). However, these processes seem to be restricted to local features and are not present when processing requires information integration from larger parts of the visual scene. This is precisely the moment when conscious perception starts to play a role, enabling the organization of distinct elements in the visual field into a coherent scene (e.g., figure-ground differentiation; Lamme, Zipser, & Spekreijse, 2002). Experimental evidence suggests that conscious vision allows for better integration of spatially or temporally distributed information, as well as higher precision of the visual representations (Ludwig, 2023). A coherent scene can then be used to guide adequate actions and predict future events. From this perspective, while object recognition is an essential part of the visual processing pipeline, it cannot fulfil the representational function of vision alone.

Another notion that complicates comparisons between humans and DNNs is temporal integration. Our perception is trained from birth on continuous perceptual input that is highly temporally correlated. Scenes are not a part of a randomized stream of unrelated snapshots. Temporal integration enables our visual system to augment the processing of stimuli with information extracted from the immediate past. This type of information can involve, for example, changes in the relative position of individuals or objects. Subsequently, this leads to one of the crucial discrepancies between human and artificial vision (the target article identifies aspects of it in sect. 4.1.1–4.1.7). DNNs are built to classify ensembles of pixels in a digital image, while human brains interpret everything as two-dimensional (2D) projections of three-dimensional (3D) objects. This fact imposes restrictions on possible interpretations of perceptual stimuli (which can lead to mistakes) but ultimately allows the visual system to not rely solely on immediate physical stimulation. This in turn makes perception more stable and useful in the context of interactions with the environment. These processes may occur without human-like consciousness. However, consciousness seems to increase the temporal integration of stimuli, strongly shaping the outcome of visual processing.

In this commentary, we aimed to justify why consciousness should be taken into account while modelling human vision with DNNs. Similar inspirations from cognitive science have proven very successful in the recent past in the case of attention (Vaswani et al., 2017) and some researchers already proposed consciousness-like mechanisms (Bengio, 2019). However, even in healthy humans, reliable measurement of consciousness is difficult both theoretically (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008) and methodologically (Wierzchoń, Paulewicz, Asanowicz, Timmermans, & Cleeremans, 2014). The task is even more challenging if one would attempt to implement such measurement in artificial neural networks (Timmermans, Schilbach, Pasquali, & Cleeremans, 2012). Nevertheless, probing the capabilities of DNNs in realizing functions connected to conscious vision might prove necessary for comparison between DNNs and humans. To make such a comparison more feasible, we would like to propose a rudimentary operationalization of subjective experience as "context dependence." In the case of visual perception, context can be defined very broadly as all the spatially or temporally distant elements of a visual scene that alter its processing. It also suggests that the global integration of perceptual features is a good approximation of the unifying function of conscious vision. Interestingly, we note that most of the phenomena mentioned in sect. 4.2 of the target article can be reformulated as examples of some form of context dependence, making this overarching principle easy to convey. Showing that DNNs are similar to humans, that is, are selectively susceptible to illusions, alter categorization based on other objects in the scene, or demonstrate object invariance, would be a strong argument in favour of the functional similarity.

Competing interest. None.

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

References

Andrillon, T., & Kouider, S. (2020). The vigilant sleeper: Neural mechanisms of sensory (de)coupling during sleep. *Current Opinion in Physiology*, *15*, 47–59. https://doi.org/10.1016/j.cophys.2019.12.002

Axelrod, V., Bar, M., & Rees, G. (2015). Exploring the unconscious using faces. *Trends in Cognitive Sciences*, *19*(1), 35–45. https://doi.org/10.1016/j.tics.2014.11.003

Bengio, Y. (2019). The consciousness prior. *arXiv*, arXiv:1709.08568. http://arxiv.org/abs/1709.08568

Fahrenfort, J. J., Snijders, T. M., Heinen, K., van Gaal, S., Scholte, H. S., & Lamme, V. A. F. (2012). Neuronal integration in visual cortex elevates face category tuning to conscious face perception. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(52), 21504–21509. https://doi.org/10.1073/pnas.1207414110

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/s1364-6613(00)01482-0

Lamme, V. A. F. (2020). Visual functions generating conscious seeing. *Frontiers in Psychology*, *11*, 83. https://doi.org/10.3389/fpsyg.2020.00083

Lamme, V. A. F., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *Journal of Cognitive Neuroscience*, *14*(7), 1044–1053. https://doi.org/10.1162/089892902320474490

Ludwig, D. (2023). The functions of consciousness in visual processing. *Neuroscience of Consciousness*, *2023*(1), niac018. https://doi.org/10.1093/nc/niac018

Mudrik, L., & Deouell, L. Y. (2022). Neuroscientific evidence for processing without awareness. *Annual Review of Neuroscience*, *45*(1), 403–423. https://doi.org/10.1146/annurev-neuro-110920-033151

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321. https://doi.org/10.1016/j.tics.2008.04.008

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: Consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1412–1423. https://doi.org/10.1098/rstb.2011.0421

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 6000–6010. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, L., Weng, X., & He, S. (2012). Perceptual grouping without awareness: Superiority of Kanizsa triangle in breaking interocular suppression. *PLoS ONE*, *7*(6), e40106. https://doi.org/10.1371/journal.pone.0040106

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27*, 109–120. https://doi.org/10.1016/j.concog.2014.04.009

Neural networks need real-world behavior

Aedan Y. Li[a] and Marieke Mur[a,b]

[a]Department of Psychology, Western University, London, ON, Canada and [b]Department of Computer Science, Western University, London, ON, Canada
aedan.li@uwo.ca, www.aedanyueli.com
mmur@uwo.ca, murlab.org

doi:10.1017/S0140525X23001504, e398

Abstract

Bowers et al. propose to use controlled behavioral experiments when evaluating deep neural networks as models of biological vision. We agree with the sentiment and draw parallels to the notion that "neuroscience needs behavior." As a promising path forward, we suggest complementing image recognition tasks with increasingly realistic and well-controlled task environments that engage real-world object recognition behavior.

Bowers et al. describe the importance of targeted behavioral experiments when evaluating deep neural networks as models of biological vision. We agree with the sentiment and draw parallels to the notion that "neuroscience needs behavior" (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). A major point raised by Bowers et al. is that one system – a neural network – can provide an excellent prediction of another system – the visual system – while relying on entirely different mechanisms. Carefully designed behavioral experiments are needed to assess how good the match really is. This point echoes the historic *multiple realizability* argument highlighted by Krakauer et al., which states that different (neural) mechanisms can solve the same computational problem. Krakauer and colleagues proposed the same solution: Carefully designed behavioral experiments, to generate and test hypotheses about the neural mechanisms that give rise to behavior. In essence, neuroscience and modeling both need behavior to guide hypothesis testing and theory development in their endeavor to understand how the brain works.

What types of behavioral experiments are best suited to evaluate deep neural networks as models of biological vision? As suggestions for the modeling community, we take inspiration from solutions pioneered by neuroscience in recent years (e.g., Snow & Culham, 2021). There is growing realization that real-world object recognition engages distinct neural responses compared to the behaviors involved with standard image recognition tasks. In the traditional experiment, observers respond with button presses to images displayed on a computer monitor as brain activity is recorded. This approach has provided important insights into biological vision and has served as a great starting point for model evaluation (e.g., Jozwik, Kietzmann, Cichy, Kriegeskorte, & Mur, 2023). However, traditional experiments do not fully capture how humans interact with objects in real-world environments.

We suggest that our experiments should increasingly mimic real-world behavior, by: (1) including tasks beyond image recognition when evaluating deep neural networks, and (2) developing platforms that enable simulation of realistic task environments. Using these environments, both humans and models can be subjected to a wide range of real-world behavioral tasks such as object tracking (e.g., following a moving animal) or visual search (e.g., finding objects in cluttered scenes); also see Peters and Kriegeskorte (2021) for discussions. The researcher will be offered a level of control that supports carefully designed experiments while maintaining ecological validity. The proposed platforms are now within reach thanks to advances in virtual reality and three-dimensional (3D) computer graphics, which are yielding powerful game engines accessible to psychologists and modelers alike. Promising recent approaches have extended the *Unity* game engine to the design of psychology experiments (e.g., Alsbury-Nealy et al., 2022; Brookes et al., 2020; Peters, Retchin, & Kriegeskorte, 2022; Starrett et al., 2021) and the simulation of interactive physics (e.g., *ThreeDWorld*; Gan et al., 2021).

Importantly, we suggest that the behavior in task environments should include the measurement of continuous dependent variables that unfold over time. Traditional cognitive psychology and neuroscience experiments use binary metrics such as "yes/no" or "multiple-choice" questions with one correct option among competitors (e.g., image classification). By contrast, humans in the real world have evolved to complete unstructured tasks in service of survival-related goals. We use cognitive abilities honed through millions of years of primate evolution and over a decade of childhood development to navigate environments, build

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

tools, find food, solve problems, and interact with other humans in cooperative and competitive settings. These dynamic behaviors involve head, body, and limb movements (Adolph & Franchak, 2017) and are based on internal decisions made from the input received from our sensory organs at millisecond timescales (Stanford, Shankar, Massoglia, Costello, & Salinas, 2010). Measuring the continuous behavioral dynamics may allow for richer understanding compared to discrete variables that average over many experimental trials (Spivey, 2007; for object memory dynamics, see Li, Yuan, Pun, & Barense, 2023; for navigation dynamics, see de Cothi et al., 2022; for "continuous psychophysics," see Straub & Rothkopf, 2022).

The models we build should also explain neural activity measured as humans complete different experimental tasks. Not only will this approach create a wealth of interdisciplinary opportunities, but modelers could take advantage of psychology and neuroscience theory which continues to make important predictions about behavior (e.g., Behrens et al., 2018; Cowell, Barense, & Sadil, 2019). As one example, the anterior temporal lobes are theorized to be a centralized "hub" region of the human brain involved in combining multiple sensory features to form object concepts (Lambon Ralph, Jefferies, Patterson, & Rogers, 2017). This structure supports the formation of new concepts in tasks involving the combination of 3D shape and sound (Li et al., 2022). Furthermore, damage to the anterior temporal lobes results in predictable impairments on memory, perception, and learning tasks (i.e., *semantic dementia*; Barense, Rogers, Bussey, Saksida, & Graham, 2010; Hodges & Patterson, 2007). A complete model should be able to make novel predictions about behavioral and brain responses while also accounting for existing data across many tasks.

We have outlined concrete suggestions toward a collaborative path that we envision to be productive. We suggest that modelers should design realistic tasks in virtual reality, measure the continuous behavioral dynamics that unfold over time, and assess correspondences to brain activity across many tasks. However, there are also many challenges that lie ahead before these suggestions can be fully realized: The expertise required to span cognitive psychology and neuroscience in addition to computational modeling is daunting. Developing naturalistic real-world experiments requires programming skills often not taught in psychology and neuroscience curriculums, whereas theoretical models important for understanding human cognition are often not taught in computer science. Fully characterizing the dynamics of behavior and brain activity will likely require theory and measurement techniques that have not yet been developed (Druckmann & Rust, 2023). For these reasons, we suggest an incremental, highly interdisciplinary and collaborative approach toward real-world experiments, which we hope will lead to a more complete understanding of how the human brain may support object-centered representations.

Our suggestions reemphasize the centrality of behavior – described as "psychological findings" by Bowers et al. – across both the development of more human-like neural networks as well as in the continued understanding of the human brain.

Financial support. A. Y. L. is supported by a BrainsCAN Postdoctoral Fellowship. M. M. is supported by an NSERC Discovery Grant.

Competing interest. None.

References

Adolph, K. E., & Franchak, J. M. (2017). The development of motor behavior. *Wiley Interdisciplinary Reviews. Cognitive Science*, *8*(1–2), e1430. https://doi.org/10.1002/wcs.1430

Alsbury-Nealy, K., Wang, H., Howarth, C., Gordienko, A., Schlichting, M. L., & Duncan, K. D. (2022). OpenMaze: An open-source toolbox for creating virtual navigation experiments. *Behavior Research Methods*, *54*, 1374–1387. https://doi.org/10.3758/s13428-021-01664-9

Barense, M. D., Rogers, T. T., Bussey, T. J., Saksida, L. M., & Graham, K. S. (2010). Influence of conceptual knowledge on visual object discrimination: Insights from semantic dementia and MTL amnesia. *Cerebral Cortex*, *20*(11), 2568–2582. https://doi.org/10.1093/cercor/bhq004

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., & Mushtaq, F. (2020). Studying human behavior with virtual reality: The unity experiment framework. *Behavior Research Methods*, *52*, 455–463. https://doi.org/10.3758/s13428-019-01242-0

Cowell, R. A., Barense, M. D., & Sadil, P. S. (2019). A roadmap for understanding memory: Decomposing cognitive processes into operations and representations. *eNeuro*, *6*(4), ENEURO.0122-19.2019. https://doi.org/10.1523/ENEURO.0122-19.2019

de Cothi, W., Nyberg, N., Griesbauer, E. M., Ghanamé, C., Zisch, F., Lefort, J. M., … Spiers, H. J. (2022). Predictive maps in rats and humans for spatial navigation. *Current Biology: CB*, *32*(17), 3676–3689.e5. https://doi.org/10.1016/j.cub.2022.06.090

Druckmann, S., & Rust, N. C. (2023). Unraveling the entangled brain: How do we go about it? *Journal of Cognitive Neuroscience*, *35*, 368–371. https://doi.org/10.1162/jocn_a_01950

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., … Yamins, D. L. K. (2021). ThreeDWorld: A platform for interactive multi-modal physical simulation. *bioRxiv*. https://doi.org/10.48550/arXiv.2007.04954

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: A unique clinicopathological syndrome. *The Lancet. Neurology*, *6*(11), 1004–1014. https://doi.org/10.1016/S1474-4422(07)70266-1

Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N., & Mur, M. (2023). Deep neural networks and visuo-semantic models explain complementary components of human ventral-stream representational dynamics. *The Journal of Neuroscience*, *43*(10), 1731–1741. https://doi.org/10.1523/JNEUROSCI.1424-22.2022

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Li, A. Y., Ladyka-Wojcik, N., Qazilbash, H., Golestani, A., Walther, D. B., Martin, C. B., & Barense, M. D. (2022). Multimodal object representations rely on integrative coding. *bioRxiv*. https://doi.org/10.1101/2022.08.31.504599

Li, A. Y., Yuan, J. Y., Pun, C., & Barense, M. D. (2023). The effect of memory load on object reconstruction: Insights from an online mouse-tracking task. *Attention, Perception & Psychophysics*, *85*(5), 1612–1630. https://doi.org/10.3758/s13414-022-02650-9

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*(9), 1127–1144. https://doi.org/10.1038/s41562-021-01194-6

Peters, B., Retchin, M., & Kriegeskorte, N. (2022). Flying objects: Challenging humans and machines in dynamic object vision. *Cognitive Computational Neuroscience*. https://doi.org/10.32470/ccn.2022.1301-0

Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*(6), 506–519. https://doi.org/10.1016/j.tics.2021.02.008

Spivey, M. (2007). *The continuity of mind*. Oxford University Press.

Stanford, T. R., Shankar, S., Massoglia, D. P., Costello, M. G., & Salinas, E. (2010). Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience*, *13*(3), 379–385. https://doi.org/10.1038/nn.2485

Starrett, M. J., McAvan, A. S., Huffman, D. J., Stokes, J. D., Kyle, C. T., … Ekstrom, A. D. (2021). Landmarks: A solution for spatial navigation and memory experiments in virtual reality. *Behavior Research Methods*, *53*, 1046–1059. https://doi.org/10.3758/s13428-020-01481-6

Straub, D., & Rothkopf, C. A. (2022). Putting perception into action with inverse optimal control for continuous psychophysics. *eLife*, *11*, e76635. https://doi.org/10.7554/eLife.76635

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

The scientific value of explanation and prediction

Hause Lin[a,b]

[a]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA and [b]Hill and Levene Schools of Business, University of Regina, Regina, SK, Canada
hauselin@gmail.com
https://www.hauselin.com

doi:10.1017/S0140525X23001735, e399

Abstract

Deep neural network models have revived long-standing debates on the value of explanation versus prediction for advancing science. Bowers et al.'s critique will not make these models go away, but it is likely to prompt new work that seeks to reconcile explanatory and predictive models, which could change how we determine what constitutes valuable scientific knowledge.

Explanatory power and predictive accuracy are different qualities, but are they inconsistent or incompatible? Bowers et al.'s critique of deep neural network models of biological vision resurfaces age-old debates and controversial questions in the history of science (Breiman, 2001; Hempel & Oppenheim, 1948). First, must an explanatory model have predictive accuracy to be considered scientifically valuable? Similarly, must a predictive model have explanatory power to have scientific value? Second, what kinds of models are better for advancing scientific knowledge, and how should we determine the scientific value of models?

To appreciate the significance of Bowers et al.'s critique, let us consider explanation and prediction as two orthogonal dimensions rather than two extremes on a continuum. As shown in Figure 1a, some of the most successful models and theories in the history of humankind have occupied different positions in this two-dimensional space: Theories like relativity and quantum electrodynamics are located in the top-right quadrant (i.e., very high explanatory power and predictive accuracy), whereas Darwinian evolution sits at the bottom-right quadrant (i.e., high explanatory power but little predictive accuracy, or at least cannot be tested for predictive accuracy yet). Importantly, successful models in disciplines ranging from physics to biology generally have high explanatory power.

Younger disciplines such as neuroscience and psychology – to which biological vision belongs – often aspire to emulate more established disciplines by developing models and theories with increasing explanatory power over time. Bowers et al. also prefer explanatory models and emphasize the importance of using controlled laboratory experimentation to test causal mechanisms and develop explanatory models and theories. Since researchers in these disciplines have historically valued models with explanatory power more than those with predictive accuracy, the consequence is that existing models are mostly located in the bottom two quadrants (Fig. 1a; some explanatory power but relatively low predictive accuracy). Models with high predictive accuracy are rare or even unheard of (e.g., Eisenberg et al., 2019; Yarkoni & Westfall, 2017).

Neural network models of biological vision have therefore introduced a class of scientific models that occupies a unique location in the two-dimensional space in Figure 1a (top-left quadrant). One could even argue that it might be the first time the discipline (including neuroscience and psychology) has produced models that have greater predictive accuracy than explanatory power. If so, it should come as no surprise that researchers – many of whom have been trained to rely primarily on experimentation to test theories – would feel uncomfortable with models with such different qualities and even question the scientific value of these models, despite recent calls to integrate explanation and prediction in neighboring disciplines (Hofman et al., 2021; Yarkoni & Westfall, 2017).

The current state of research on deep neural network models of biological vision reflects a critical juncture in the history of neuroscience as well as psychological and social science.

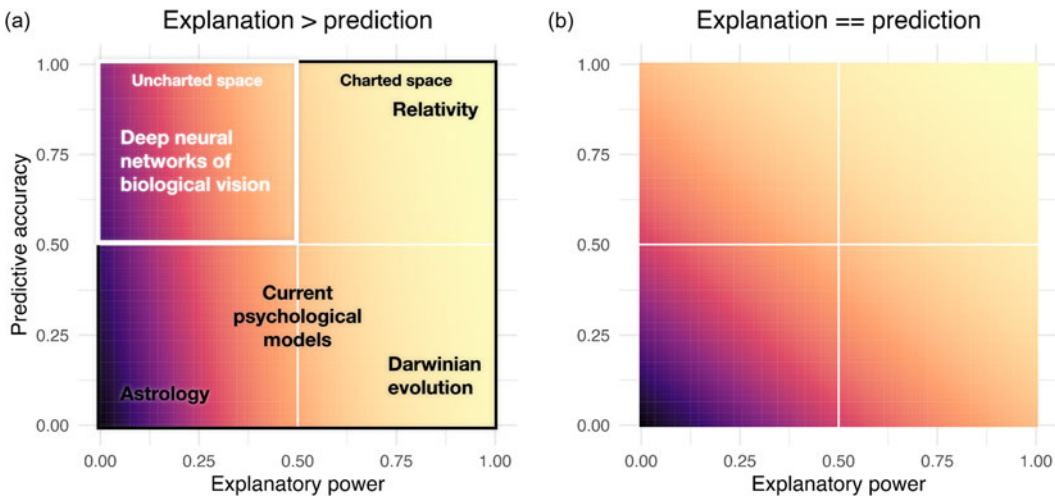


Figure 1 (Lin). Scientific value of models with different degrees of two qualities: Explanatory power and predictive accuracy. (a) Bowers et al. value explanation over prediction, such that models with greater explanatory power are preferred. (b) Alternative value function that values both qualities equally. Hotter colors denote greater scientific value, whereas cooler colors denote less scientific value.

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

The long-standing tension between different philosophical approaches to theory development no longer exists only in the abstract – arguably for the first time, researchers have to reconcile, in practice, explanatory models with their predictive counterparts.

Bowers et al. emphasize the value of experimentation and the need for models to explain a wide range of experimental results. But this approach is not without limitations: When experiments and models become overly wedded to each other, models might lose touch with reality because they explain phenomena only within but not beyond the laboratory (Lin, Werner, & Inzlicht, 2021).

Should explanation be favored over prediction? The prevailing approach to theory development has certainly favored explanation (Fig. 1a), but the state of research on deep neural network models suggests that developing models with predictive accuracy might be a complementary approach that could help to test the relevance of explanatory models that have been developed through controlled experimentation. Predictive models could also be used to discover new explanations or causal mechanisms. If so, it is conceivable that current and future generations of researchers (who have been trained to also consider predictive accuracy) might come to value explanation and prediction equally (Fig. 1b).

Deep neural network models are becoming increasingly popular in a wide range of academic disciplines. Although Bowers et al.'s critique is unlikely to reverse this trend, it highlights how new methods and technological advances can turn age-old philosophical debates into practical issues researchers now have to grapple with. How the explanatory and predictive approaches are reconciled or integrated in the coming years by researchers working on biological vision is likely to have far-reaching consequences on how researchers in other disciplines think about theory development and the philosophy of science. And it is also likely to reshape our views of what constitutes valid and valuable scientific knowledge.

Acknowledgments. I thank Adam Bear and Alexandra Decker for helpful discussions.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-10301-1

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., … Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science*, *16*(4), 854–863. https://doi.org/10.1177/1745691620974773

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Fixing the problems of deep neural networks will require better training data and learning algorithms

Drew Linsley and Thomas Serre

Department of Cognitive Linguistic & Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI, USA
drew_linsley@brown.edu
thomas_serre@brown.edu
https://sites.brown.edu/drewlinsley
https://serre-lab.clps.brown.edu

doi:10.1017/S0140525X23001589, e400

Abstract

Bowers et al. argue that deep neural networks (DNNs) are poor models of biological vision because they often learn to rival human accuracy by relying on strategies that differ markedly from those of humans. We show that this problem is worsening as DNNs are becoming larger-scale and increasingly more accurate, and prescribe methods for building DNNs that can reliably model biological vision.

Over the past decade, vision scientists have turned to deep neural networks (DNNs) to model biological vision. The popularity of DNNs comes from their ability to achieve human-level performance on visual tasks (Geirhos et al., 2021) and the seemingly concomitant correspondence of their hidden units with biological vision (Yamins et al., 2014). Bowers et al. marshal evidence from psychology and neuroscience to argue that while DNNs and biological systems may achieve similar accuracy on visual benchmarks, they often do so by relying on qualitatively different visual features and strategies (Baker, Lu, Erlikhman, & Kellman, 2018; Malhotra, Evans, & Bowers, 2020, 2022). Based on these findings, Bowers et al. call for a reevaluation of what DNNs can tell us about biological vision and suggest dramatic adjustments going forward, potentially even moving on from DNNs altogether. Are DNNs the wrong paradigm for modeling biological vision?

Systematically evaluating DNNs for biological vision

While this commentary identifies multiple shortcuts in DNNs that are commonly used in vision science, such as ResNet and AlexNet, it does not delve into the root causes of these issues or how widespread they are across different DNN architectures and training routines. We previously addressed these questions with *ClickMe*, a web-based game in which human participants teach DNNs how to recognize objects by highlighting category-diagnostic visual features (Linsley, Eberhardt, Sharma, Gupta, & Serre, 2017; Linsley, Shiebler, Eberhardt, & Serre, 2019). With *ClickMe*, we collected annotations of the visual features that humans rely on to recognize approximately 25% of ImageNet images (https://serre-lab.github.io/Harmonization/). Human feature importance maps from *ClickMe* reveal startling regularity: Animals were categorized by their faces, whereas inanimate objects like cars were categorized by their wheels and headlights

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

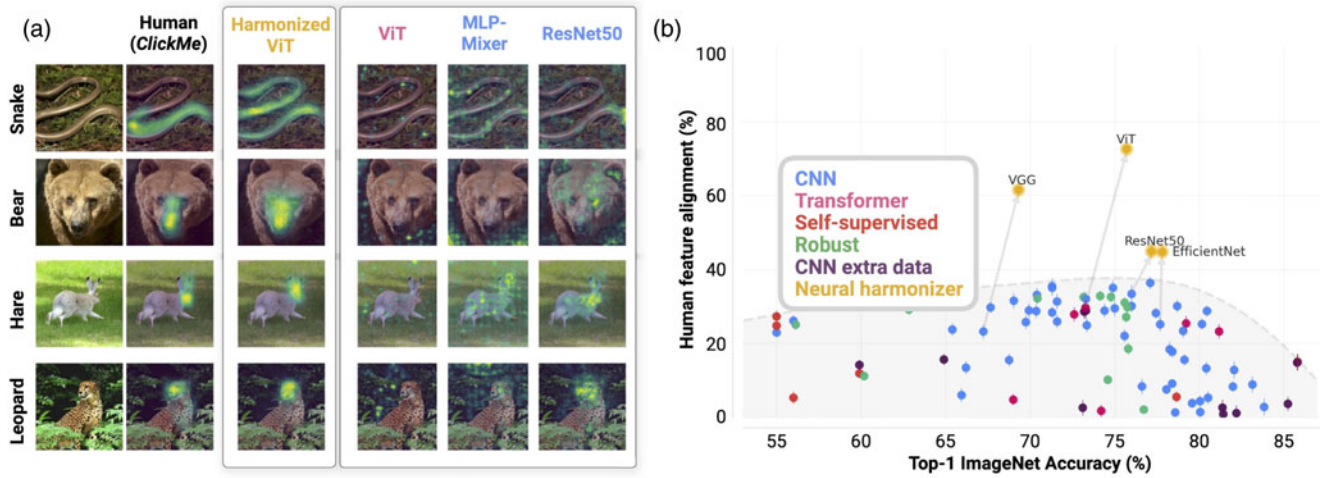


Figure 1 (Linsley and Serre). A growing misalignment between biological vision and DNNs (adapted from Fel et al., 2022). (a) Diagnostic features for object classification differ between humans and DNNs. (b) The Spearman correlation between human and DNN feature importance maps is decreasing as a function of DNN accuracy on ImageNet. This trade-off can be addressed with the neural harmonizer — a method for explicitly aligning DNN representations with humans for object recognition.

(Fig. 1a). Human participants were also significantly more accurate at rapid object classification when basing their decisions on these features rather than image saliency. In contrast, while DNNs sometimes selected the same diagnostic features as humans, they often relied on "shortcuts" for object recognition (Geirhos et al., 2020). For example, a DNN called the Vision transformer (ViT) relied on background features, like grass, to recognize a hare, whereas human participants focused almost exclusively on the animal's head (Fig. 1a). Even more concerning is that the visual features and strategies of humans and DNNs are becoming increasingly misaligned as newer DNNs become more accurate (Fig. 1b). We and others have observed similar trade-offs between DNN accuracy on ImageNet and their ability to explain various human behavioral data and psychophysics (Fel, Felipe, Linsley, & Serre, 2022; Kumar, Houlsby, Kalchbrenner, & Cubuk, 2022). Our work indicates that the mismatch between DNN and biological vision identified by Bowers et al. is pervasive and worsening.

The next generation of DNNs for biological vision

Bowers et al. argue that the inability of DNNs to learn human-like visual strategies reflects architectural limitations. They are correct that there is a rich literature demonstrating how mechanisms inspired by neuroscience can improve the capabilities of DNNs, helping them learn perceptual grouping (Kim, Linsley, Thakkar, & Serre, 2020; Linsley, Kim, Ashok, & Serre, 2019a; Linsley, Kim, Veerabadran, Windolf, & Serre, 2018, 2021), visual reasoning (Kim, Ricci, & Serre, 2018; Vaishnav et al., 2022; Vaishnav & Serre, 2023), robust object recognition (Dapello et al., 2020), and to more accurately predict neural activity (Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021; Kubilius et al., 2018; Nayebi et al., 2018). The other fundamental difference between DNNs and biological organisms is how they learn; humans and DNNs learn from vastly different types of data with presumably different objective functions. We believe that the limitations raised by Bowers et al. result from a mismatch in data diets and objective functions because we were able to significantly improved the alignment of DNNs with humans by introducing *ClickMe* data into their training routines ("Neural harmonizer," Fig. 1).

Biologically inspired data diets and objective functions

We believe that the power of DNNs for biological vision is from their ability to generate computational- and algorithmic-level hypotheses about vision, which will guide experiments to identify plausible circuits. For instance, the great success of gradient descent and backpropagation for training DNNs has inspired the search for biologically plausible approximations (Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020). Visual neuroscience is similarly positioned to benefit from DNNs if we can improve their alignment with biology.

The most straightforward opportunity for aligning DNNs with biological vision is to train them with more biologically plausible data and objective functions (Smith & Slone, 2017 Richards et al., 2019). There have been efforts to do this with first-person video, however, these efforts have failed to yield much benefit in computer vision or other aspects of biological vision (Orhan, Gupta, & Lake, 2020; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021; Zhuang et al., 2021), potentially because the small scale of these datasets makes them ill-suited for training DNNs. An alternative approach is to utilize advances in three-dimensional (3D) computer vision, like neural radiance fields (Mildenhall et al., 2020), to generate spatiotemporal (and stereo) datasets for training DNNs that are infinitely scalable and can be integrated with other modalities, such as somatosensation and language. It is also very likely that objective functions that will lead to human-like visual strategies and features from these datasets have yet to be discovered. However, promising directions include optimizing for slow feature analysis (Wiskott & Sejnowski, 2002) and predictive coding (Lotter, Kreiman, & Cox, 2016; Mineault, Bakhtiari, Richards, & Pack, 2021), which could help align DNNs with humans without relying on *ClickMe* data.

Aligned DNNs may be all we need

Bowers et al. point out a number of ways in which DNNs fail as models of biological vision. These problems are pervasive and likely caused by the standard image datasets and training routines of DNNs, which are guided by engineering rather than biology. Bowers et al. may well be right that an entirely new class of

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

models is needed to account for biological vision, but at the moment there are no viable alternatives. Until other model classes can rival human performance on visual tasks, we suspect that the most productive path forward toward modeling biological vision and aligning DNNs with biological vision is to develop more biologically plausible data diets and objective functions.

Acknowledgments. We thank Lakshmi Govindarajan for his helpful comments and feedback on drafts of this commentary.

Financial support. This work was supported by ONR (N00014-19-1-2029), NSF (IIS-1912280), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004).

Competing interest. None.

References

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.

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178.

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. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 13073–13087). Curran.

Fel, T., Felipe, I., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 9432–9446). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/3d681cc4487b97c08e5aa67224dd74f2-Paper-Conference.pdf

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., 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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 23885–23899). Curran.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2020). Disentangling neural mechanisms for perceptual grouping. In Z. Chen, J. Zhang, M. Arjovsky, & L. Bottou (Eds.), *International Conference on Learning Representations*, Addis Abada, Ethopia.

Kim, J., Ricci, M., & Serre, T. (2018). Not-So-CLEVR: Learning same-different relations strains feedforward neural networks. *Interface Focus*, *8*(4), 20180011.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, 408385. https://doi.org/10.1101/408385

Kumar, M., Houlsby, N., Kalchbrenner, N., & Cubuk, E. D. (2022). Do better ImageNet classifiers assess perceptual similarity better? https://openreview.net › forumhttps://openreview.net › forum. https://openreview.net/pdf?id=qrGKGZZvH0

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews. Neuroscience*, *21*(6), 335–346.

Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., & Serre, T. (2017). What are the visual features underlying human versus machine vision? In Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, & M. Yang (Eds.), *IEEE international conference on computer vision workshops,* Venice, Italy (pp. 2706–2714).

Linsley, D., Kim, J., Ashok, A., & Serre, T. (2019a). Recurrent neural circuits for contour detection. *International conference on representation learning*. https://openreview.net/forum?id=H1gB4RVKvB&noteId=H1gB4RVKvB

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31, pp. 152–164). Curran.

Linsley, D., Malik, G., Kim, J., Govindarajan, L. N., Mingolla, E., & Serre, T. (2021). Tracking without re-recognition in humans and machines. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 19473–19486). Curran.

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2019). Learning what and where to attend. In I. Loshchilov & F. Hutter (Eds.), *7th International conference on representation learning*, New Orleans.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv [cs.LG]*. http://arxiv.org/abs/1605.08104

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

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.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. *arXiv [cs.CV]*. http://arxiv.org/abs/2003.08934

Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 28757–28771). Curran.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., … Yamins, D. L. K. (2018). Task-driven convolutional recurrent models of the visual system. *arXiv [q-bio.NC]*. http://arxiv.org/abs/1807.00053

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 9960–9971). Curran.

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.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, *8*, 2124.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind: Discoveries in Cognitive Science*, *5*, 20–29.

Vaishnav, M., Cadene, R., Alamia, A., Linsley, D., VanRullen, R., & Serre, T. (2022). Understanding the computational demands underlying visual reasoning. *Neural Computation*, *34*(5), 1075–1099.

Vaishnav, M., & Serre, T. (2023). GAMR: A guided attention model for (visual) reasoning. *International conference on learning representations*. https://openreview.net/pdf?id=iLMgk2IGNyv

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770.

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

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118

The model-resistant richness of human visual experience

Jianghao Liu[a,b] and Paolo Bartolomeo[a]

[a]Sorbonne Université, Inserm, CNRS, Paris Brain Institute, ICM, Hôpital de la Pitié-Salpêtrière, Paris, France and [b]Dassault Systèmes, Vélizy-Villacoublay, France

jianghao.liu@icm-institute.org

paolo.bartolomeo@icm-institute.org

doi:10.1017/S0140525X23001656, e401

Abstract

Current deep neural networks (DNNs) are far from being able to model the rich landscape of human visual experience. Beyond visual recognition, we explore the neural substrates of visual mental imagery and other visual experiences. Rather than shared visual representations, temporal dynamics and functional connectivity of the process are essential. Generative adversarial networks may drive future developments in simulating human visual experience.

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

Bowers et al. report several lines of evidence challenging the alleged similarities between deep neural network (DNN) models of visual recognition and their biological counterparts. However, human visual experience is not limited to visual recognition. In addition to the case of visual illusion presented by Bowers et al., it is important for models of the human visual system to consider a range of other visual experiences, including visual hallucinations, dreams, and mental imagery. For example, most of us can "visualize" objects in their absence, by engaging in visual mental imagery. Using partially shared neural machinery used for visual perception, visual mental imagery allows us to make predictions based on past experiences, imagine future possibilities, and simulate the possible outcomes of our decisions. Our commentary focuses on these relationships and is structured into four key points.

First, shared neural substrates of visual perception and visual mental imagery include high-level visual regions in the ventral temporal cortex (Bartolomeo, Hajhajate, Liu, & Spagna, 2020; Spagna, Hajhajate, Liu, & Bartolomeo, 2021). In the absence of visual input, these regions are activated top-down by other systems, such as the semantic system and the frontoparietal attention networks. Bowers et al. highlighted the challenge of modeling top-down activity with feedforward DNNs. It is currently believed that the visual system relies on distinct feedback signals to cortical layers and exhibits individual temporal dynamics for different visual experiences. In particular, visual stimulation modulates activities in mid-layers, while contextual information or illusory content feedbacks to superficial layers, and visual imagery feedbacks to deeper cortical layers (Bergmann, Morgan, & Muckli, 2019; Muckli et al., 2015). Visual imagery exhibits temporal overlap with perceptual processing during late stages of processing (Dijkstra, Mostert, Lange, Bosch, & van Gerven, 2018), likely corresponding to activity in the ventral temporal cortex but not in the early visual cortex (Spagna et al., 2021). In contrast, patients with Charles Bonnet hallucinations show a gradual increase in activity in the early visual cortex, which then gradually decreases as it moves further along the visual hierarchy (Hahamy, Wilf, Rosin, Behrmann, & Malach, 2021).

Second, evidence from neuropsychology, neuroimaging, and direct cortical stimulation suggests striking differences in the activity of the ventral temporal cortex in the two hemispheres when processing visual information (Liu, Spagna, & Bartolomeo, 2022b). While direct cortical electrical stimulation tends to produce visual hallucinatory experiences predominantly when applied to the right temporal lobe, there is a strong lateralization to the left hemisphere for voluntary visual mental imagery. These asymmetries could potentially stem from particular hemispheric networks' predispositions toward constructing mental models of the external environment or verifying them through real-world testing (Bartolomeo & Seidel Malkinson, 2022). After unilateral brain strokes, in some cases the healthy hemisphere can compensate for the visual deficit (Bartolomeo & Thiebaut de Schotten, 2016). At present, DNN models do not incorporate either hemispheric asymmetries or the potential reorganization of these asymmetries following a stroke.

Third, some otherwise neurotypical individuals show unusually weak or strong visual mental imagery (aphantasia and hyperphantasia) (Keogh, Pearson, & Zeman, 2021; Milton et al., 2021). Aphantasic individuals perform visual imagery and visual perceptual tasks with similar accuracy than typical imagers, but with slower response times (Liu & Bartolomeo, 2023). Consistent with these behavioral results, ultra-high field fMRI shows similar activation patterns between typical imagers and individuals with congenital aphantasia (Liu et al., 2023). The fusiform imagery node, a high-level visual region in the left-hemisphere ventral temporal cortex (Spagna et al., 2021), coactivates with dorsolateral frontoparietal networks in typical imagers, but is functionally isolated from these networks in aphantasic individuals during both imagery and perception. These findings suggest that high-level visual information in the ventral cortical stream is not sufficient to generate a conscious visual experience, and that a functional disconnection from frontoparietal networks may be responsible for the lack of experiential content in visual mental imagery in aphantasic individuals.

Fourth, in line with the previous point on the importance of frontoparietal networks, the way we subjectively experience both perceptions and mental images relies heavily on the interaction with other cognitive processes, such as attention and visual working memory. Despite their importance, these factors are not taken into account in DNN modeling. A recent study using human intracerebral recordings and single-layer recurrent neural network modeling found that the dynamic interactions between specific frontoparietal attentional networks and high-level visual areas play a crucial role in conscious visual perception (Liu et al., 2023).

This evidence from the biological human brain can inspire future developments of DNNs in simulating the cognitive architecture of human visual experience. Generative adversarial networks may be promising candidates to drive these efforts forward. For instance, imagery mechanisms could act as the generator of quasi-perceptual experiences, while reality monitoring could serve as the discriminator to distinguish between sensory inputs from real or imagined sources (Gershman, 2019; Lau, 2019). Recent studies investigated involuntary visual experiences using generative neural network models, such as in memory replay (van de Ven, Siegelmann, & Tolias, 2020), intrusive imagery (Cushing et al., 2023), and adversarial dreaming (Deperrois, Petrovici, Senn, & Jordan, 2022). Regarding voluntary visual mental imagery, some key strategies may involve modeling the retrieval process of representations pertaining to semantic information and visual features (Liu et al., 2023), and incorporating biologically inspired recurrence in visual imagery processing (Lindsay, Mrsic-Flogel, & Sahani, 2022).

In conclusion, we suggest that shared representations in visual cortex are not the primary factor in generating and distinguishing distinct visual experiences. Rather, the temporal dynamics and functional connectivity of the process are essential. Current DNNs are inadequate to accurately model the complexity of human visual experience. Biologically inspired generative adversarial networks may provide novel ways of simulating the varieties of human visual experience.

Financial support. J. L. received funding from Dassault Systèmes. The work of P. B. is supported by the Agence Nationale de la Recherche through ANR-16-CE37-0005 and ANR-10-IAIHU-06, and by the Fondation pour la Recherche sur les AVC through FR-AVC-017.

Competing interest. None.

References

Bartolomeo, P., Hajhajate, D., Liu, J., & Spagna, A. (2020). Assessing the causal role of early visual areas in visual mental imagery. *Nature Reviews Neuroscience*, *21*(9), 517. https://doi.org/10.1038/s41583-020-0348-5

Bartolomeo, P., Seidel Malkinson, T. (2022). Building models, testing models: Asymmetric roles of SLF III networks?: Comment on "Left and right temporal-parietal

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

junctions (TPJs) as 'match/mismatch' hedonic machines: A unifying account of TPJ function" by Doricchi et al. *Physics of Life Reviews*, *44*, 70–72. https://doi.org/10/grrsd8
Bartolomeo, P., & Thiebaut de Schotten, M. (2016). Let thy left brain know what thy right brain doeth: Inter-hemispheric compensation of functional deficits after brain damage. *Neuropsychologia*, *93*, 407–412. https://doi.org/10/f9g9wb
Bergmann, J., Morgan, A. T., & Muckli, L. (2019). Two distinct feedback codes in V1 for "real" and "imaginary: Internal experiences. *bioRxiv*, 664870. https://doi.org/10.1101/664870
Cushing, C. A., Dawes, A. J., Hofmann, S. G., Lau, H., LeDoux, J. E., & Taschereau-Dumouchel, V. (2023). A generative adversarial model of intrusive imagery in the human brain. *PNAS Nexus*, *2*(1), pgac265. https://doi.org/10.1093/pnasnexus/pgac265
Deperrois, N., Petrovici, M. A., Senn, W., & Jordan, J. (2022). Learning cortical representations through perturbed and adversarial dreaming. *eLife*, *11*, e76384. https://doi.org/10.7554/eLife.76384
Dijkstra, N., Mostert, P., Lange, F. P., Bosch, S., & van Gerven, M. A. (2018). Differential temporal dynamics during visual imagery and perception. *eLife*, *7*, e33904. doi: https://doi.org/10.7554/eLife.33904
Gershman, S. J. (2019). The generative adversarial brain. *Frontiers in Artificial Intelligence*, *2*, 486362. doi: https://www.frontiersin.org/articles/10.3389/frai.2019.00018
Hahamy, A., Wilf, M., Rosin, B., Behrmann, M., & Malach, R. (2021). How do the blind "see"? The role of spontaneous brain activity in self-generated perception. *Brain*, *144*(1), 340–353. https://doi.org/10.1093/brain/awaa384
Keogh, R., Pearson, J., & Zeman, A. (2021). Aphantasia: The science of visual imagery extremes. *Handbook of Clinical Neurology*, *178*, 277–296. https://doi.org/10.1016/B978-0-12-821377-3.00012-X
Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. *PsyArXiv*. https://doi.org/10.31234/osf.io/ckbyf
Lindsay, G. W., Mrsic-Flogel, T. D., & Sahani, M. (2022). Bio-inspired neural networks implement different recurrent visual processing strategies than task-trained ones do. *bioRxiv*, 2022.03.07.483196. https://doi.org/10.1101/2022.03.07.483196
Liu, J., & Bartolomeo, P. (2023). Probing the unimaginable: The impact of aphantasia on distinct domains of visual mental imagery and visual perception. *Cortex*, *166*, 338–347. doi: 10.1016/j.cortex.2023.06.003
Liu, J., Bayle, D. J., Spagna, A., Sitt, J. D., Bourgeois, A., Lehongre, K., … Bartolomeo, P. (2023). Fronto-parietal networks shape human conscious report through attention gain and reorienting. *Communications Biology*, *6*, 730. doi: 10.1038/s42003-023-05108-2
Liu, J., Spagna, A., & Bartolomeo, P. (2022b). Hemispheric asymmetries in visual mental imagery. *Brain Structure and Function*, *227*(2), 697–708. https://doi.org/10.1007/s00429-021-02277-w
Liu, J., Zhan, M., Hajhajate, D., Spagna, A., Dehaene, S., Cohen, L., & Bartolomeo, P. (2023). Ultra-high field fMRI of visual mental imagery in typical imagers and aphantasic individuals. *bioRxiv*. https://doi.org/10.1101/2023.06.14.544909
Milton, F., Fulford, J., Dance, C., Gaddum, J., Heuerman-Williamson, B., Jones, K., … Zeman, A. (2021). Behavioral and neural signatures of visual imagery vividness extremes: Aphantasia versus hyperphantasia. *Cerebral Cortex Communications*, *2*(2), tgab035. https://doi.org/10.1093/texcom/tgab035
Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., … Yacoub, E. (2015). Contextual feedback to superficial layers of V1. *Current Biology*, *25*(20), 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057
Spagna, A., Hajhajate, D., Liu, J., & Bartolomeo, P. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neuroscience & Biobehavioral Reviews*, *122*, 201–217. https://doi.org/10.1016/j.neubiorev.2020.12.029
van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17866-2

You can't play 20 questions with nature and win redux

Bradley C. Love[a] and Robert M. Mok[b]

[a]Experimental Psychology, University College London, London, UK and [b]MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK
b.love@ucl.ac.uk
Rob.Mok@mrc-cbu.cam.ac.uk
https://bradlove.org/
https://sites.google.com/site/robmokbrainbob/

doi:10.1017/S0140525X23001747, e402

Abstract

An incomplete science begets imperfect models. Nevertheless, the target article advocates for jettisoning deep-learning models with some competency in object recognition for toy models evaluated against a checklist of laboratory findings; an approach which evokes Alan Newell's 20 questions critique. We believe their approach risks incoherency and neglects the most basic test; can the model perform its intended task.

The first author remembers a discussion with fellow graduate students in the late 1990s. Each offered a prediction for when a model would be able to take photographic images as inputs and provide labels. Predictions ranged from a hundred years into the future to never. Similar estimates were provided for speech recognition. Albeit imperfect, we now have models that can perform both tasks. We marvel at the speed of progress and how poorly placed cognitive scientists were to anticipate it. In fairness, perhaps it would take that long to achieve these results if models were built by psychologists on the basis of their laboratory studies! Related discussions may have occurred in adjacent fields, such as linguistics.

In their target article, the authors correctly note some limitations of deep networks as models of vision. However, every model in an incomplete science is imperfect so these criticisms are largely benign, especially in a field that is rapidly progressing. The authors' key critique seems to be that image-computable models (i.e., models that actually attempt object recognition) are poor models of human vision because they do not account for findings from a selected set of laboratory studies. The authors invite us to return to the halcyon days before deep learning to a time of box-and-arrow models in cognitive psychology and "blocks world" models of language (Winograd, 1971), when modelers could narrowly apply toy models to toy problems safe in the knowledge that they would not be called upon to generalize beyond their confines nor pave the way for future progress.

Essentially, the authors are advocating for what Alan Newell cautioned against in his classic essay, "You can't play 20 questions with nature and win" (Newell, 1973). Newell worried that all the clever experiments psychologists conducted would not integrate into any coherent understanding of cognition. We agree – it seems unlikely progress will be made by amassing yet more laboratory findings. What will tie all these results together to make them more than cognitive science trivia?

One answer is models. Perhaps the most basic test for a model is whether it can perform its intended task. Once the model has some basic competency, then secondary questions can be considered, like how well the model accounts for aspects of human behavior and brain response. A model that cannot pass the first hurdle, such as an object recognition model that cannot process sensory inputs (e.g., photographic images), is of little use for understanding how the brain accomplishes such feats. Models that can apply to the task can be compared on how well they account for human data (i.e., model selection). Completing the scientific loop, competing models can guide empirical efforts by suggesting informative experiments that tease apart their predictions. Instead, the authors advocate for skipping the crucial step of considering models that have basic competency and proceeding to evaluating accounts against a checklist of selected findings from laboratory studies.

This 20 questions mindset naturally pairs with the falsification approach the authors advocate. However, we do not share their

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

enthusiasm for falsifying models that are a priori wrong and incomplete. Instead, we suggest a Bayesian or evidential philosophy of science is more appropriate in which one aims for the model that is most likely given the data (which could include data from laboratory studies). Of course, the most important empirical finding to address for a model of human object recognition is basic competency in object recognition. It seems odd to worry about fine-grain distinctions observed in the laboratory studies when the basics are missing; it is like worrying about a car's window tint when it lacks an engine and transmission.

Finally, the authors seem oddly reluctant to acknowledge or engage with work that successfully addresses their criticisms. For example, they criticize correlative approaches to assessing correspondences between brain regions and models layers, such representation similarity analyses (RSAs) and encoding approaches, but neglect to mention work that has successfully addressed these deficiencies. Recent work evaluates correspondences under the mantra "correlation does imply correspondence" by directly interfacing brain activity with a model layer to evaluate whether brain activity can drive the model toward an appropriate output (i.e., behavior; Sexton & Love, 2022). Notice this approach requires a model that can perform object recognition, which further highlights the value of image-computable models in evaluating neurocomputational hypotheses. Another example is the authors' omission of large-scale "prediction" studies that successfully identify deficiencies in deep-learning models and adjudicate between competing models. For example, Roads and Love (2021) derived an embedding of 50k images based on human judgments and found all deep-learning models diverged from human semantic judgments with better performing models from an engineering perspective being less human aligned. This type of large-scale study provides a general and stringent test of how human aligned representations are in deep-learning models. The authors mention that deep networks are susceptible to shortcut learning, which is true, but they neglect to discuss the literature devoted to ameliorating this issue, including approaches that successfully address the authors' own manipulations (e.g., adding a colored dot to an image) to create shortcuts (Dagaev et al., 2023). The authors state that comparing models differing on a single factor is uncommon despite such comparisons being standard in machine learning papers, referred to as ablation studies. All these cases indicate that progress is being made with image-computable models on the very issues the authors highlight.

In conclusion, the fact that deep networks with some competency in object recognition fail to account for findings from some laboratory tasks has led the authors to conclude deep-learning models are of limited value. One might instead conclude that the laboratory studies themselves are limited in paving the way toward a complete model of human vision. After all, our preconceived notions of how vision works guide these study designs. Some laboratory studies will prove fundamental to explaining human vision, some will be irrelevant. It seems to us that one will never be able to determine which is which in the absence of models with basic competencies.

Financial support. This work was supported by ESRC (ES/W007347/1), Wellcome Trust (WT106931MA), and a Royal Society Wolfson Fellowship (18302) to B. C. L., and the Medical Research Council UK (MC UU 00030/7) and a Leverhulme Trust Early Career Fellowship (Leverhulme Trust, Isaac Newton Trust: SUAI/053 G100773, SUAI/056 G105620, ECF-2019-110) to R. M. M.

Competing interest. None.

References

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., & Love, B. C. (2023). A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, *166*, 164–171. https://doi.org/10.1016/j.patrec.2022.12.010

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.). (1973). Visual information processing: Proceedings of the 8th annual Carnegie symposium on cognition, held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972. Academic Press.

Roads, B. D., & Love, B. C. (2021). Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3547–3557).

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*, 1–9. doi: https://doi.org/www.science.org/doi/10.1126/sciadv.abm2219

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. *AITR*-235. http://hdl.handle.net/1721.1/7095

Explananda and *explanantia* in deep neural network models of neurological network functions

Mihnea Moldoveanu

Desautels Centre for Integrative Thinking, Rotman School of Management, University of Toronto, Toronto, ON, Canada
mihnea.moldoveanu@rotman.utoronto.ca
https://www.rotman.utoronto.ca/FacultyAndResearch/Faculty/FacultyBios/Moldoveanu

doi:10.1017/S0140525X23001632, e403

Abstract

Depending on what we mean by "explanation," challenges to the explanatory depth and reach of deep neural network models of visual and other forms of intelligent behavior may need revisions to both the elementary building blocks of neural nets (the explananda) and to the ways in which experimental environments and training protocols are engineered (the explanantia). The two paths assume and imply sharply different conceptions of how an explanation explains and of the explanatory function of models.

By one definition of "explanation," enjoining deep neural network researchers to directly address the functional capabilities and behaviors exhibited by the neurological networks of human brains – as Bowers et al. do – is similar to requiring engineers of digital clocks to use their to explain the specific workings of analog watch subassemblies and components. It also highlights the differences between *simulation* and *emulation* on the one hand, and between *explanation* and *prediction* or *enaction* on the other. Each of these approaches spells out a very different conception of what explanations are for, and how they explain.

A digital and an analog clock both keep (relative) time. An analog clock does so by converting energy stored in the translational movement of a spring into the rotational energy of a set of gears whose ratios are designed to register time lapses indexed by the movement of a catchment that indexes a set duration (of, say, 1 second). A digital clock starts with a crystal oscillator vibrating at a high frequency (e.g., 60 Hz), which is digitally subdivided

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

down to the desired frequency (1 Hz) of a wave whose consecutive peaks mark the desired interval. Although the inner components and assemblies of the clocks are very different, the digital clock "models" the analog clock in the sense that it replicates its function; and, vice versa. But the two clocks do not share common components, which is why we cannot use our understanding of the digital clock's mechanisms to understand "how the analog clock works," if by *understand* we mean *emulate*, or, replicate at the component and subassembly level the function of each component or subassembly of an entity – in the way in which one digital computer can *emulate* the workings of another by executing all of the functions of each of its assemblies.

Explanation as means for emulation: *The "artificial replication" approach*. If we take the "emulation" route to explanation, we are confronted with a component-level "modeling mismatch" between neurons in a neural network storing information in weights that are integer or rational (i.e., finite-precision) numbers and neurons in biological neurons whose weights can be real numbers that are theoretically capable of storing infinite amounts of information, and, even if truncated, their resolution can be adaptively varied (Balcazar, Gavalda, & Siegelmann, 1997). This mismatch cannot be offset by creating neural nets that merely mimic the heterogeneity of human neurons and the topology of brain networks to test for relationships between structure and function "one at a time," even if we model a single neuron by a deep neural net (Beniaguev, Segev, & London, 2021): There is a degree of freedom (weight quantization) that is missing from the model. Moreover, DNNs work in discrete and fixed time steps and do not therefore adequately replicate the fluidly adaptive time constants of real neurological assemblies. And, the smoothing nonlinearities artificially introduced in neural networks to satisfy regularity and convergence properties are introduced ad hoc, to optimize for the properties of an output, rather than allowed to emerge and evolve as a function of time.

So, if by "understanding" we mean that *explanantia* need to be *emulated* by the *explananda*, then we need to engineer building blocks for deep neural networks that heed the continuity and adaptive informational breadth of neurological networks. One example is the design of liquid time constant networks (Hasani, Lechner, Amini, Rus, & Grosu, 2020), built from assemblies of linear, first-order dynamical systems connected by nonlinear gates, which embody dynamical systems with variable ("liquid") time constants, and achieve higher levels of expressiveness (while maintaining stability) than do their counterparts with fixed time steps and hard-wired nonlinearities. One can alternatively seek to relax the constraint on quantization or resolution for the weights of a neural network (Jia, Lam, & Althoefer, 2022) to more closely resemble the features of their cortical counterparts.

Explanation as means to prediction and production: The "invisible hand approach". On the contrary, we can take the view that *all and only* what "understanding" an entity means is *predicting* and *producing* the behaviors it exhibits. This is the approach deep neural net designers have taken, at the cost of abstracting away well-defined tasks such as object classification and time series prediction from the panoply of human capabilities, to engineer simple reward functions. Plausibly, this simplificatory approach to neural net engineering has contributed to the divergence of the fields of visual neuroscience and automatic pattern recognition and image classification that Bowers et al. point to. It is, then, unsurprising that deep neural networks currently in use do not replicate human functions such as combinatorial generation of the "possible ways an object can look when turned" and the parsing of two-dimensional (2D) scenes for depth reconstruction and whole-part decomposition: Learning to perform these tasks requires different – and often more complicated – reward functions, that track the multiplicity of ways in which a human uses vision in the wild and the multiplicity of goals one might have when "looking." Introducing a human in the training loop of a machine is equivalent to creating rewards that encode the complex credit assignment map of a task (designing successful communicative acts) without having to specify, *ex ante*, why or how that complexity arises. Tellingly, the recent advances in the performance of large language models (e.g., GPT2 to GPT3.5 via InstructGPT) are traceable not only to the increase in the parameter space of the new models, but, more importantly, to the use "human-in-the-loop" reinforcement learning (Ouyang et al., 2022) that incorporates feedback from untrained humans that do not "understand the underlying model" but answer questions (e.g., "Helpful? Honest? Harmless?") in ways that help fine tune it in accordance with a set of human preferences over sequences of acts that induce a multi-dimensional objective function ("what is a successful communicative act?") which the raters may *also* not fully understand. One does not have to "know what one is doing" to sufficiently "understand" an environment from sparse inputs provided by people who also do not explicitly "know what they are doing" to provide them.

Financial support. This work was supported by the Desautels Centre for Integrative Thinking, Rotman School of Management, University of Toronto.

Competing interest. None.

References

Balcazar, J. C., Gavalda, R., & Siegelmann, H. T. (1997). Computational power of neural networks: A characterization in terms of Kolmogorov complexity. *IEEE Transactions on Information Theory*, *43*, 1175–1183.

Beniaguev, D., Segev, I., & London, M. (2021) Single cortical neurons as deep artificial neural networks. *Neuron*, *109*, 2727–2739.

Hasani, R., Lechner, M., Amini, A., Rus, D., & Grosu, R. (2020). Liquid time constant networks. *arXiv:2006.04439v4*.

Jia, G., Lam, H.-K., & Althoefer, K. (2022). Variable weight algorithm for convolutional neural networks and its applications to classification of seizure phases and types. *Pattern Recognition*, *121*, 108226.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv: 2203.02155v1*.

Going after the bigger picture: Using high-capacity models to understand mind and brain

Hans Op de Beeck[a] and Stefania Bracci[b]

[a]Leuven Brain Institute, KU Leuven, Leuven, Belgium and [b]Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy
hans.opdebeeck@kuleuven.be
stefania.bracci@unitn.it
www.hoplab.be
https://webapps.unitn.it/du/en/Persona/PER0076943/Curriculum

doi:10.1017/S0140525X2300153X, e404

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

Abstract

Deep neural networks (DNNs) provide a unique opportunity to move towards a generic modelling framework in psychology. The high representational capacity of these models combined with the possibility for further extensions has already allowed us to investigate the forest, namely the complex landscape of representations and processes that underlie human cognition, without forgetting about the trees, which include individual psychological phenomena.

Bowers et al. challenge the notion that deep neural networks (DNNs) are the best or even a highly promising model of human cognition and recommend that future studies should test specific psychological neural phenomena and potential hypotheses by independently manipulating factors.

We agree with Bowers et al. that overall predictive power is not sufficient to have a good model, in particular not when experiments are lacking in diversity of tested stimuli (Grootswagers & Robinson, 2021). Nevertheless, prediction is a necessary condition and a good starting point. DNNs have the power to serve as a generic model, and at the same time they can be tested on a variety of cognitive/psychological phenomena to go beyond prediction and give insight to understand the functioning of a system. Strikingly, and in contrast to how the literature is characterized by Bowers et al., the first wave of studies comparing DNNs to human vision already included studies that went beyond mere prediction on generic stimulus sets. To just take one example, the study by Kubilius, Bracci, and Op de Beeck (2016) that is characterized as a prediction-based experiment by Bowers et al., tested a specific cognitive hypothesis (the role of nonaccidental properties, see Biederman, 1987) and independently manipulated shape and category similarity (Kubilius et al., 2016; see also Bracci & Op de Beeck, 2016; Zeman, Ritchie, Bracci, & Op de Beeck, 2020). More in general, the goal of explanation over prediction is already a central one as shown by examples of recent work testing underlying mechanisms of object perception (Singer, Seeliger, Kietzmann, & Hebart, 2022), category domains (Dobs, Martinez, Kell, & Kanwisher, 2022), or predictive coding (Ali, Ahmad, de Groot, van Gerven, & Kietzmann, 2022), just to mention a few. The wealth of data that the community has gathered with DNNs in less than a decade illustrates the potential of this approach.

Bowers et al. provide many examples of failures of DNNs, on the side admitting some of the successes and progress. Many of the failures show that vanilla DNNs, as is true for all models, are not perfect and do not capture all aspects of brain processing. Revealing such limitations is generally considered essential to move the field forward towards making DNN computations more human-like (Firestone, 2020), and is no reason to abandon these models as long as there is an obvious road ahead with them. Some proposed examples are the addition of optical limitations reminiscent of the human eye that can make a network more robust to adversarial attacks (Elsayed et al., 2018), the implementation of intuitive physics (Piloto, Weinstein, Battaglia, & Botvinick, 2022), or considerations about the influence of visual system maturation and low visual acuity at birth (Avberšek, Zeman, & Op de Beeck, 2021; Jinsi, Henderson, & Tarr, 2023).

It is difficult to reconcile a fundamental criticism of DNNs that they do not capture all psychological phenomena without further extensions, with the proposal of Bower et al to switch to alternative strategies that are much more limited in terms of the extent to which they capture the full complexity of information processing from input to output (e.g., Grossberg, 1987; Hummel & Biederman, 1992; McClelland, Rumelhart, & PDP Research Group, 1986, Psych Rev). These alternative models are very appealing but also more narrow in scope. Consider, for example, the simplicity with which the well-known ALCOVE model explains categorization (Kruschke, 1992), compared to the complex high-dimensional space that is the actual reality of the underlying representations (for a review, see Bracci & Op de Beeck, 2023). Note that we consider these alternatives to be an excellent way to obtain a conceptual understanding of a phenomenon, we all very much build on top of this pioneering work using conceptually elegant models with few parameters (e.g., Ritchie & Op de Beeck, 2019). Nevertheless, scientists should not stop there. If we would, then we would be left with a wide range of niche solutions and no progress towards either a generic model that can be applied across domains, or at least a path towards it. Luckily, this path looks very promising for DNNs, given that there is a large community of relatively junior scientists that is ready to make progress (e.g., Doerig et al., 2023; Naselaris et al., 2018). The necessary modifications will move the needle in various directions, such as elaborations in terms of front-ends, architecture, learning and optimization rules, learning regime, level of neural detail (e.g., spiking networks), the addition of attentional and working memory processes, and potentially the interaction with symbolic processing. None of that will lead to the dismissal of DNNs.

We see the high capacity of DNNs as a feature, not a bug, and currently we are still on the part of the curve where higher capacity means better (Elmoznino & Bonner, 2022). In contrast to the alternatives, DNNs confront us upfront with the complexity of human information processing because they have to work vis-à-vis an actual stimulus as an input. This is not just a faits divers, it is a necessary condition for the ideal model. DNNs and related artificial intelligence (AI) models seem to be able to stand up to this challenge, even up to the point that in some domains they can already predict empirical data about neural selectivity to real images to a greater extent than professors in cognitive neuroscience (Ratan Murty, Bashivan, Abate, DiCarlo, & Kanwisher, 2021). The general applicability of these models and the legacy of knowledge that has by now been obtained provides a unique resource to test a wide variety of psychological and neural phenomena (e.g., Duyck, Bracci, & Op de Beeck, 2022; Kanwisher, Gupta, & Dobs, 2023).

The way forward is to build better models, including DNN-based models that take the complexity of human vision and cognition seriously (Bracci & Op de Beeck, 2023). As it has been since the very early days of AI, we need continuous interaction and exchange between disciplines and their expertise at all levels (cognitive and computational psychologists, computer vision scientists, philosophers of the mind, neuroscientists) to bring us towards a common goal of a human-like AI that we understand mechanistically. Solving the deep problem of understanding biological vision will not happen by too easily dismissing DNNs and missing out on their potential.

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

Financial support. H. O. B. is supported by FWO research project G073122N and KU Leuven project IDN/21/010.

Competing interest. None.

References

Ali, A., Ahmad, N., de Groot, E., van Gerven, M. A. J., & Kietzmann, T. C. (2022). Predictive coding is a consequence of energy efficiency in recurrent neural networks. *Patterns*, *3*(12), 100639.

Avberšek, L. K., Zeman, A., & Op de Beeck, H. (2021). Training for object recognition with increasing spatial frequency: A comparison of deep learning with human vision. *Journal of Vision*, *21*(10), 14–14.

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

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience*, *36*(2), 432–444.

Bracci, S., & Op de Beeck, H. P. (2023). Understanding human object vision: A picture is worth a thousand representations. *Annual Review of Psychology*, *74*, 113–135.

Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, *8*(11), eabl8913.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., … Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450.

Duyck, S., Bracci, S., & Op de Beeck, H. (2022). A computational understanding of zoomorphic perception in the human brain. *bioRxiv*, 2022-09.

Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality. *bioRxiv*, 2022-07.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in Neural Information Processing Systems*, *31*.

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

Grootswagers, T., & Robinson, A. K. (2021). Overfitting the literature to one set of stimuli and data. *Frontiers in Human Neuroscience*, *15*, 682661.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, *11*(1), 23–63.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480.

Jinsi, O., Henderson, M. M., & Tarr, M. J. (2023). Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *PLoS ONE*, *18* (1), e0280145.

Kanwisher, N., Gupta, P., & Dobs, K. (2023). CNNs reveal the computational implausibility of the expertise hypothesis. *iScience*, 105976.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22.

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.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing* (Vol. 2, pp. 20–21). MIT Press.

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., … Kay, K. (2018). Cognitive computational neuroscience: A new conference for an emerging discipline. *Trends in Cognitive Sciences*, *22*(5), 365–367.

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267.

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), 5540.

Ritchie, J. B., & Op de Beeck, H. (2019). A varying role for abstraction in models of category learning constructed from neural representations in early visual cortex. *Journal of Cognitive Neuroscience*, *31*(1), 155–173.

Singer, J. J., Seeliger, K., Kietzmann, T. C., & Hebart, M. N. (2022). From photos to sketches – How humans and deep neural networks process objects across different levels of visual abstraction. *Journal of Vision*, *22*(2), 4–4.

Zeman, A. A., Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific Reports*, *10*(1), 2453.

Models of vision need some action

Constantin Rothkopf[a,b,c,d], Frank Bremmer[c,d,e], Katja Fiehler[c,d,f], Katharina Dobs[c,d,f] and Jochen Triesch[b,c,d]

[a]Centre for Cognitive Science, Technical University of Darmstadt, Darmstadt, Germany; [b]Frankfurt Institute for Advanced Studies, Goethe-Universität Frankfurt, Frankfurt am Main, Germany; [c]Center for Mind, Brain and Behavior, University of Marburg and Justus Liebig University Giessen, Giessen, Germany; [d]HMWK-Clusterproject The Adaptive Mind, Hesse, Germany; [e]Applied Physics and Neurophysics, University of Marburg, Marburg, Germany and [f]Experimental Psychology, Justus Liebig University Giessen, Giessen, Germany
constantin.rothkopf@cogsci.tu-darmstadt.de
frank.bremmer@physik.uni-marburg.de
Katja.Fiehler@psychol.uni-giessen.de
katharina.dobs@psychol.uni-giessen.de
triesch@fias.uni-frankfurt.de
https://www.theadaptivemind.de/

doi:10.1017/S0140525X23001577, e405

Abstract

Bowers et al. focus their criticisms on research that compares behavioral and brain data from the ventral stream with a class of deep neural networks for object recognition. While they are right to identify issues with current benchmarking research programs, they overlook a much more fundamental limitation of this literature: Disregarding the importance of action and interaction for perception.

Computationally, perception, cognition, and action are inseparably intertwined in sequential, goal-directed behavior (Kessler, Frankenstein, & Rothkopf, 2022). However, the branch of research considered in Bowers et al. focuses on a single visual task, that of assigning single, discrete labels of object identity to images. This is as if the whole goal of human vision was to learn to shout out an appropriate word while being presented a random pile of photographs. But, in the words of Thomas H. Huxley, the nineteenth-century English biologist and anthropologist: "The great end of life is not knowledge but action." Perception is not l'art-pour-l'art. Instead, it occurs continuously in space and time as we perform structured tasks in a complex and dynamic environment (Fiehler & Karimpur, 2023). Perception guides action and action, in turn, impacts perception (Bremmer, Churan, & Lappe, 2017; Bremmer & Krekelberg, 2003; Eckmann, Klimmasch, Shi, & Triesch, 2020; Fiehler, Brenner, & Spering, 2019). Without action, we could not make changes in the world or interact with others. Here we argue that many of the limitations of current deep neural networks (DNNs) pointed out by Bowers et al. are likely rooted in a flawed and limited framing of perception and implausible supervised learning objectives, that recent DNNs represent fruitful avenues for overcoming some of these limitations, but that we must extend current models to account for the different functions of vision: Perception, cognition, and action and how they interact. Acknowledging that perception and action are intimately related has fundamental consequences. Here we highlight five key consequences.

The sensory input to biological visual systems is highly structured as it unfolds during goal-directed behavior. Accordingly, DNNs should be trained not on independent images presented in random order with corresponding labels, but in self-supervised ways by observing continuous, structured datasets, that is, events unfolding in space and time. Many real-world objects, such as animals or faces, are not just static entities, but move dynamically and nonrigidly (Dobs, Bülthoff, & Schultz, 2018). One potential avenue currently being explored is using forms of time-based self-supervised deep learning (Orhan, Gupta, & Lake, 2020; Schneider, Xu, Ernst, Yu, & Triesch, 2021; Zhuang et al., 2021), which form invariant object representations by mapping sequences of views onto close-by latent representations without the need for labels. These models also have the potential to capture dynamic aspects of object recognition, such as the perception of dynamic faces, which cannot be captured well by current models trained on static images (Jiahui et al., 2022).

The structure of sensory input is in large part dependent on the observer's own actions. Thus, object perception and vision in general can only be understood in the context of an active, exploratory, multi-sensory observer, a view also reflected in current experimental work (Ayzenberg & Behrmann, 2023). Supervised approaches miss the impact of goal-directed action and interaction on structuring visual representations (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). Accordingly, models should learn in self-supervised ways while interacting with their environment. Indeed, visual representations have been shown to be dependent on the active visual policy (Rothkopf, Weisswange, & Triesch, 2009). Going beyond pure self-supervised invariance learning, a recent approach considers the benefits of active control of the view point for learning object representations (Xu & Triesch, 2023). Mimicking visual input from self-generated object manipulations, it learns a hierarchical representation to satisfy the two complementary desiderata of being partly invariant to viewpoint changes while at the same time permitting to predict which action is responsible for a particular change in the representation.

Learning and adaptation must be a continuous process, not limited to discrete training and test phases, but occurring continually during extended interactions with the environment. Recent approaches involving DNNs have addressed the challenge of continual learning (Wang, Liu, Duan, Kong, & Tao, 2022). However, the breadth of the required continuous adaptation to changing conditions (Roelfsema & Holtmaat, 2018; Schmitt et al., 2021) and the delicate balance of the classic stability–plasticity dilemma are still open problems for current DNNs.

The learning objectives must permit rich and adaptive representations that can feed multiple forms of interacting with the world. Instead, many of the studies considered by Bowers et al. relate to the single task of object recognition simply because the vast majority of current DNN approaches to vision select a task that gets away with ignoring actions: Attaching labels to images. Few current NN models conceptualize visual tasks in terms of visual routines, with some exceptions applying the framework of reinforcement learning to sequential visual behaviors (Araslanov, Rothkopf, & Roth, 2019). Promising directions are to jointly investigate a broad range of visual tasks (Dwivedi, Bonner, Cichy, & Roig, 2021) and to investigate those computational visual tasks relevant for action, which are predominantly attributed to the dorsal stream, and considering ecologically relevant cost functions that can account for dorsal stream properties in the primate brain (Mineault, Bakhtiari, Richards, & Pack, 2021).

Models will need to properly compute the interactions of sensory uncertainties, internally model uncertain beliefs, and the action variabilities to successfully achieve the organism's goals in sequential, adaptive behavior. Bowers et al. do not mention uncertainty once in their article. Current DNN models are not well suited to the computations required for proper belief propagation in sequential perception and action under uncertainty as required in extended behavior, where they are inseparably intertwined. As an example, humans use their perception and their actions actively to shape their internal beliefs about landmarks in navigation (Kessler et al., 2022). In their critique, Bowers et al. ignore the major computational challenge, which requires making accurate causal inferences about the origins of uncertainty in sensory data and adaptive motor output (Straub & Rothkopf, 2022).

In conclusion, we agree with Bowers et al.'s critique, but if we want to fully understand human vision including object recognition, our models must embrace the fact that vision is intimately intertwined with action in behaving, goal-directed agents.

Financial support. The research reported herein was supported by the "The Adaptive Mind," funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art.

Competing interest. None.

References

Araslanov, N., Rothkopf, C. A., & Roth, S. (2019). Actor-critic instance segmentation. In L. Davis, P. Torr, & S.-Z. Zhu (Eds.), Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, California, 16–20 June 2019 (pp. 8237–8246).

Ayzenberg, V., & Behrmann, M. (2023). The where, what, and how of object recognition. *Trends in Cognitive Sciences*, *27*, 335–336.

Bremmer, F., Churan, J., & Lappe, M. (2017). Heading representations in primates are compressed by saccades. *Nature Communications*, *8*, 920.

Bremmer, F., & Krekelberg, B. (2003). Seeing and acting at the same time: Challenges for brain (and) research. *Neuron*, *38*, 367–370.

Dobs, K., Bülthoff, I., & Schultz, J. (2018). Use and usefulness of dynamic face stimuli for face perception studies – A review of behavioral findings and methodology. *Frontiers in Psychology*, *9*, 1355.

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, *17*(8), e1009267.

Eckmann, S., Klimmasch, L., Shi, B. E., & Triesch, J. (2020). Active efficient coding explains the development of binocular vision and its failure in amblyopia. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 6156–6162.

Fiehler, K., Brenner, E., & Spering, M. (2019). Prediction in goal-directed action. *Journal of Vision*, *19*(9), 10, 1–21.

Fiehler, K., & Karimpur, H. (2023). Spatial coding for action across spatial scales. *Nature Reviews Psychology*, *2*, 72–84.

Jiahui, G., Feilong, M., di Oleggio Castello, M. V., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2022). Modeling naturalistic face processing in humans with deep convolutional neural networks. *bioRxiv*, 1–39.

Kessler, F., Frankenstein, J., & Rothkopf, C. A. (2022). A dynamic Bayesian actor model explains endpoint variability in homing tasks. *bioRxiv*, 1–25.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*, 480–490.

Mineault, P. J., Bakhtiari, S., Richards, B. A., & Pack, C. C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, *34*, 28757–28771.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*, 9960–9971.

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, *19*, 166–180.

Rothkopf, C. A., Weisswange, T. H., & Triesch, J. (2009). Learning independent causes in natural images explains the space variant oblique effect. In M. Amine, N. Enayati, & H.

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

Li (Eds.), 2009 IEEE 8th international conference on development and learning, Shanghai, China, 5–7 June 2009 (pp. 1–6). IEEE.
Schmitt, C., Schwenk, J. C. B., Schütz, A., Churan, J., Kaminiarz, A., & Bremmer, F. (2021). Preattentive processing of visually guided self-motion in humans and monkeys. *Progress in Neurobiology*, *205*, 102117.
Schneider, F., Xu, X., Ernst, M. R., Yu, Z., & Triesch, J. (2021). Contrastive learning through time. In SVRHM 2021 Workshop@NeurIPS.
Straub, D., & Rothkopf, C. A. (2022). Putting perception into action with inverse optimal control for continuous psychophysics. *eLife*, *11*, 76635.
Wang, Z., Liu, L., Duan, Y., Kong, Y., & Tao, D. (2022). Continual learning with lifelong vision transformer. In R. Chellappa, J. Matas, L. Quan, & M. Shah (Eds.), Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, Louisiana, 19–24 June 2022 (pp. 171–181).
Xu, X., & Triesch, J. (2023). CIPER: Combining invariant and equivariant representations using contrastive and predictive learning. http://arxiv.org/abs/2302.02330
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 of the United States of America*, *118*(3), e2014196118.

Perceptual learning in humans: An active, top-down-guided process

Heleen A. Slagter

Department of Cognitive Psychology, Institute for Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
h.a.slagter@vu.nl
https://research.vu.nl/en/persons/heleen-slagter

doi:10.1017/S0140525X23001644, e406

Abstract

Deep neural network (DNN) models of human-like vision are typically built by feeding blank slate DNN visual images as training data. However, the literature on human perception and perceptual learning suggests that developing DNNs that truly model human vision requires a shift in approach in which perception is not treated as a largely bottom-up process, but as an active, top-down-guided process.

Bowers et al. do the field a service with their thought-provoking commentary. If the problems currently characterizing deep neural network (DNN) models of human vision laid out in the target article are not adequately addressed, the field risks another winter. Bowers et al. sketch one important way forward: Building DNNs that can account for psychological data. I put forward that developing DNNs of human vision will foremost require a conceptual shift: From approaching perception as the outcome of a largely stimulus-driven process of feature detection and object recognition to treating perception as an active, top-down-guided process.

Building on the traditional notion of perception as a largely bottom-up process, mainstream computational cognitive neuroscience currently embraces the idea that simply feeding blank slate DNNs large amounts of training data will produce human-like vision. Yet, as Bowers et al.'s overview shows, this may not yet be the case. Based on the literature on human perceptual learning and action-oriented theories of perception, I contend that this may be a direct result from the manner in which DNNs are typically trained to "perceive": In a passive, data-driven manner. This approach typically does not induce perceptual learning that generalizes to new stimuli or tasks in humans (Lu & Dosher, 2022), is very different from how babies learn to perceive (Emberson, 2017; Zaadnoordijk, Besold, & Cusack, 2022), and does not take into account the action-oriented nature of perception (Friston, 2009; Gibson, 2014; Hurley, 2001).

The most consistent finding in the literature on visual perceptual learning in human adults is that learning is highly specific to the trained stimuli and tasks (Lu & Dosher, 2022). For example, improvements are often not observed if the test stimulus has a different orientation or contrast than the trained stimulus or when the trained stimulus is relocated or rotated (Fahle, 2004; Fiorentini & Berardi, 1980). These findings indicate that the typical outside-in approach used in perceptual learning studies in which participants are presented with stimuli to detect or categorize tends to induce learning at too low levels in the processing hierarchy to support feature-, stimulus-, or view-independent learning. Indeed, more recent research suggests that transfer learning can be enhanced when learning can be top-down guided and connect to higher levels in the processing hierarchy (Tan, Wang, Sasaki, & Watanabe, 2019). For example, when the training procedure allowed for more abstract rule formation, complete transfer of learning between physically different stimuli was observed (Wang et al., 2016). These observations fit with recent findings that perceptual learning involves higher cognitive areas (Shibata, Sagi, & Watanabe, 2014; Zhang et al., 2010) and proposals that perceptual learning is a top-down-guided process (Ahissar & Hochstein, 2004). Perceptual development in infants is also more top-down guided than traditionally assumed (Emberson, 2017) and perception continues to develop through childhood based on acquired knowledge across a range of tasks (Milne et al., 2022). Yet the building of models of human vision still typically starts from the notion that human-like vision will simply arise by feeding blank slate DNNs many supervised training images, which may cause learning at too low levels in the processing hierarchy. Indeed, as Bowers et al. summarize, DNNs can be fooled by additive noise (Heaven, 2019), have difficulty generalizing learning to novel objects, and do not form transformation-tolerant object identity representations at higher layers (Xu & Vaziri-Pashkam, 2021). These problems conceivably reflect insufficient top-down-guided learning.

Research also shows that perception and action are interdependent processes, in particular during the development of perception (Zaadnoordijk et al., 2022). For example, kittens that are passively moved around, do not develop depth perception (Held & Hein, 1963), just like DNN's that are fed visual input do not perceive depth (Jacob, Pramod, Katti, & Arun, 2021). Humans are not passive perceivers, but continuously build on past experiences to actively predict and generate their own sensory information through action, thereby top-down driving their own learning (Boonstra & Slagter, 2019; Buzsáki, 2019; Friston, 2009; Gibson, 1988). That perception incorporates expectations about the sensory outcome of actions is demonstrated by the fact that humans who wear goggles that flip the visual field from left to right, do not perceive a normal world (albeit flipped left-right), but experience distorted perception (Kohler, 1963), caused by the disruption of normal sensorimotor contingencies. Moreover, recent studies show that responses in early visual cortex also reflect actions (Schneider, 2020). These findings cannot easily be explained by the classical view of the brain as processing information serially from sensory to cognitive to motor control stages (Hurley, 2001), each subserved by distinct brain regions – a view that currently still drives much of DNN research. Rather, they indicate that perception emerges from dynamic feedback relations between

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

input and output, and does not merely entail the encoding of environmental statistics, but also the statistics of agent–environment interactions (Friston, 2010). Yet DNNs are generally trained in a passive way. This may cause shortcut learning or DNNs to latch onto features that do not matter to humans in categorizing objects. DNNs may focus on texture (Geirhos et al., 2022) or local rather than global shape (Baker, Lu, Erlikhman, & Kellman, 2018), because they never had to interact with objects, for which global shape knowledge is important. Notably, the development of global shape representations may depend on the dorsal stream (Ayzenberg & Behrmann, 2022).

To develop models of human-like vision, the field thus needs to turn the notion of perception on its head: From bottom-up driven to top-down guided and fundamentally serving agent–environment interactions. Important steps are already taken in this direction. For example, DNN architectures wired to top-down infer their sensory input has been shown to work at scale (Millidge, Salvatori, Song, Bogacz, & Lukasiewicz, 2022). There are also exciting developments in robotics, in which artificial systems equipped with the possibility to predict and generate their sensory information through action can top-down drive their own learning (Lanillos et al., 2021). DNNs have the potential to provide powerful ways to study the human brain and behavior, but this will require the incorporation of biologically realistic, action-oriented learning algorithms, grounding vision on interactions with the environment.

Financial support. H. A. S. is supported by a consolidator ERC grant "PlasticityOfMind" (101002584).

Competing interest. None.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. https://doi.org/10.1016/j.tics.2004.08.011

Ayzenberg, V., & Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*, *26*(12), 1119–1132. https://doi.org/10.1016/j.tics.2022.09.019

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. https://doi.org/10.1371/journal.pcbi.1006613

Boonstra, E. A., & Slagter, H. A. (2019). The dialectics of free energy minimization. *Frontiers in Systems Neuroscience*, *13*, 42. https://doi.org/10.3389/fnsys.2019.00042

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.

Emberson, L. L. (2017). Chapter One – How does experience shape early development? Considering the role of top-down mechanisms. In J. B. Benson (Ed.), *Advances in child development and behavior* (Vol. 52, pp. 1–41). JAI. https://doi.org/10.1016/bs.acdb.2016.10.001

Fahle, M. (2004). Perceptual learning: A case for early selection. *Journal of Vision*, *4*(10), 4. https://doi.org/10.1167/4.10.4

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature*, *287*(5777), 43–44. https://doi.org/10.1038/287043a0

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2022). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv, arXiv:1811.12231*. https://doi.org/10.48550/arXiv.1811.12231

Gibson, E. J. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. Annual *Review* of *Psychology*, *39*, 1–41.

Gibson, J. J. (2014). *The ecological approach to visual perception* (1st ed.). Routledge.

Heaven, D. (2019). Why deep-learning AIs are so easy to fool. *Nature*, *574*(7777), 163–166. https://doi.org/10.1038/d41586-019-03013-5

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, *56*, 872–876. https://doi.org/10.1037/h0040546

Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, *129*(1), 3–40. https://doi.org/10.1023/A:1012643006930

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), Article 1. https://doi.org/10.1038/s41467-021-22078-3

Kohler, I. (1963). The formation and transformation of the perceptual world. *Psychological Issues*, *3*, 1–173.

Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., … Tani, J. (2021). Active inference in robotics and artificial agents: Survey and challenges. *arXiv, arXiv:2112.01871*. https://doi.org/10.48550/arXiv.2112.01871

Lu, Z.-L., & Dosher, B. A. (2022). Current directions in visual perceptual learning. *Nature Reviews Psychology*, *1*(11), Article 11. https://doi.org/10.1038/s44159-022-00107-2

Millidge, B., Salvatori, T., Song, Y., Bogacz, R., & Lukasiewicz, T. (2022). Predictive coding: Towards a future of deep learning beyond backpropagation? *arXiv, arXiv:2202.09467*. https://doi.org/10.48550/arXiv.2202.09467

Milne, G. A., Lisi, M., McLean, A., Zheng, R., Groen, I. I. A., & Dekker, T. M. (2022). Emergence of perceptual reorganisation from prior knowledge in human development and convolutional neural networks. *BioRxiv*. https://doi.org/10.1101/2022.11.21.517321

Schneider, D. M. (2020). Reflections of action in sensory cortex. *Current Opinion in Neurobiology*, *64*, 53–59. https://doi.org/10.1016/j.conb.2020.02.004

Shibata, K., Sagi, D., & Watanabe, T. (2014). Two-stage model in perceptual learning: Toward a unified theory. *Annals of the New York Academy of Sciences*, *1316*(1), 18–28. https://doi.org/10.1111/nyas.12419

Tan, Q., Wang, Z., Sasaki, Y., & Watanabe, T. (2019). Category-induced transfer of visual perceptual learning. *Current Biology*, *29*(8), 1374–1378. e3. https://doi.org/10.1016/j.cub.2019.03.003

Wang, R., Wang, J., Zhang, J.-Y., Xie, X.-Y., Yang, Y.-X., Luo, S.-H., … Li, W. (2016). Perceptual learning at a conceptual level. *The Journal of Neuroscience*, *36*(7), 2238–2246. https://doi.org/10.1523/JNEUROSCI.2732-15.2016

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-22244-7

Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*(6), 510–520. https://doi.org/10.1038/s42256-022-00488-2

Zhang, J.-Y., Zhang, G.-L., Xiao, L.-Q., Klein, S. A., Levi, D. M., & Yu, C. (2010). Rule-based learning explains visual perceptual learning and its specificity and transfer. *The Journal of Neuroscience*, *30*(37), 12323–12328. https://doi.org/10.1523/JNEUROSCI.0704-10.2010

Comprehensive assessment methods are key to progress in deep learning

Michael W. Spratling

Department of Informatics, King's College London, London, UK
michael.spratling@kcl.ac.uk
https://nms.kcl.ac.uk/michael.spratling/

doi:10.1017/S0140525X23001668, e407

Abstract

Bowers et al. eloquently describe issues with current deep neural network (DNN) models of vision, claiming that there are deficits both with the methods of assessment, and with the models themselves. I am in agreement with both these claims, but propose a different recipe to the one outlined in the target article for overcoming these issues.

The target article proposes that deep neural networks (DNNs) be assessed using controlled experiments that evaluate changes in

model behaviour as all but one variable is kept constant. Such experiments might provide information about the similarities and differences between brains and DNNs, and hence, spur development of DNNs better able to model the biological visual system. However, in reality work in deep learning is concerned with developing methods that work, irrespective of the biological plausibility of those methods: Deep learning is an engineering endeavour driven by the desire to produce DNNs that perform the "best." Even in the subdomain where brain-like behaviour is a consideration (Schrimpf et al., 2020) the desire is to produce DNNs that produce the best performance. Hence, if controlled experiments were introduced, the results would almost certainly be summarised by a single value so that the performance of competing models could be ranked, and as a consequence there would be little to distinguish these new experimental methods from current ones.

What is meant by "best" performance, and how is it assessed, is the key issue. While training samples and supervision play a role in deep learning analogous to nurture during brain development, assessment plays a role analogous to that of evolution: Determining which DNNs are seen as successful, and hence, which will become the basis for future research efforts. The evaluation methods accepted as standard by a research community thus have a huge influence on progress in that field. Different evaluation methods might be adopted by different fields, for example classification accuracy on unseen test data might be accepted in computer vision, while Brain-Score or the sort of controlled experiments advocated by the target article might be used to evaluate models of biological vision. However, as is comprehensively catalogued in the target article, current DNNs suffer from such a range of severe defects that they are clearly inadequate either as models of vision or as reliable methods for computer vision. Both research agendas would, therefore, benefit from more rigorous and comprehensive evaluation methods that can adequately gauge progress.

Given the gross deficits of current DNNs, it seems premature to assess them in fine detail against psychological and neurobiological data. Rather, their performance should be evaluated by testing the ability to generalise to changes in viewing conditions (Hendrycks & Dietterich, 2019; Michaelis et al., 2019; Mu & Gilmer, 2019; Shen et al., 2021), the ability to reject samples from categories that were not seen during training (Hendrycks & Gimpel, 2017; Vaze, Han, Vedaldi, & Zisserman, 2022), the ability to reject exemplars that are unlike images of any object (Kumano, Kera, & Yamasaki, 2022; Nguyen, Yosinski, & Clune, 2015), and robustness to adversarial attacks (Biggio & Roli, 2018; Croce & Hein, 2020; Szegedy et al., 2014).

Methods already exist for testing generalisation and robustness of this type; the problem is that they are not routinely used, or that models are assessed using one benchmark but not others. The latter is particularly problematic, as there are likely to be trade-offs between performance on different tasks. The trade-off between adversarial robustness and clean accuracy is well known (Tsipras, Santurkar, Engstrom, Turner, & Madry, 2019), but others are also likely to exist. For example, improving the ability to reject unknown classes is likely to reduce performance on classifying novel samples from known classes, as such exemplars are more likely to be incorrectly seen as unknown. Hence, efforts to develop a model that is less deficient in one respect, may be entirely wasted as the resulting model may be more deficient in another respect. Only when the community routinely requires comprehensive evaluation of models for generalisation and robustness will progress be made in reducing the range of deficits exhibited by models. Once such progress has been made it will be necessary to expand the range of assessments performed in order to effectively distinguish the performance of competing models and to spur further progress to address other deficiencies. The range of assessments might eventually be expanded to include neurophysiological and psychophysical tests.

The assessment regime advocated here can only be applied to models that are capable of processing images, and hence, would not be applicable to many models proposed in the psychology and neuroscience literatures. The target article advocates expanding assessment methods to allow such models to be evaluated and compared to DNNs. However, the ability to process images would seem to me to be a minimum requirement for a model of vision, and models that cannot be scaled to deal with images are not worth evaluating.

To perform well in terms of generalisation and robustness it seems likely that DNNs will require new mechanisms. As Bowers et al. say, it is unclear if suitable mechanisms can be learnt purely from the data. Indeed, even a model trained on 400 million images fails to generalise well (Radford et al., 2021). The target article also points out that biological visual systems do not need to learn many abilities (such as adversarial robustness, tolerance to viewpoint, etc.), and instead these abilities seem to be "built-in." Brains contain many inductive biases: The nature side of the nature–nurture cooperation that underlies brain development. These biases underlie innate abilities and behaviours (Malhotra, Dujmović, & Bowers, 2022; Zador, 2019) and constrain and guide learning (Johnson, 1999; Zaadnoordijk, Besold, & Cusack, 2022). Hence, as advocated in the target article, and elsewhere (Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Malhotra, Evans, & Bowers, 2020; Zador, 2019), biological insights can potentially inspire new mechanisms that will improve deep learning. However, work in deep learning does not need to be restricted to only considering inductive biases that are biologically inspired, especially as there are currently no suggestions as to how to implement many potentially useful mechanisms which humans appear to use. Indeed, if better models of biological vision are to be developed it is essential that work in neuroscience and psychology contribute useful insights. Unfortunately, the vast majority of such work so far has concentrated on cataloguing "where" and "when" events happen (where an event might be a physical action, neural spiking, fMRI activity, etc.). Such information is of no use to modellers who need information about "how" and "why."

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317–331. doi:10.1016/j.patcog.2018.07.023

Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In H. Daumé III & A. Singh (Eds.), *Proceedings of the international conference on machine learning, volume 119 of Proceedings of machine learning research* (pp. 2206–2216). arXiv:2003.01690.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*, 245–258. doi:10.1016/j.neuron.2017.06.011

Hendrycks, D., & Dietterich, T. G. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the international conference on learning representations*, New Orleans, USA. arXiv:1903.12261.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the international conference on Learning representations*, Toulon, France. arXiv:1610.02136.

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

Johnson, M. H. (1999). Ontogenetic constraints on neural and behavioral plasticity: Evidence from imprinting and face recognition. *Canadian Journal of Experimental Psychology*, *53*, 77–90.
Kumano, S., Kera, H., & Yamasaki, T. (2022). Are DNNs fooled by extremely unrecognizable images? *arXiv*, arXiv:2012.03843.
Malhotra, G., Dujmović, M., & Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modelling visual object recognition. *PLoS Computational Biology*, *18*(5), e1009572. doi:10.1371/journal.pcbi.1009572
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. doi:10.1016/j.visres.2020.04.013
Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., … Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv*, arXiv:1907.07484.
Mu, N., & Gilmer, J. (2019). MNIST-C: A robustness benchmark for computer vision. *arXiv*, arXiv:1906.02337.
Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv*, arXiv:1412.1897.
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*, arXiv:2103.00020. https://proceedings.mlr.press/v139/radford21a.html
Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423 https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X
Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2021). Towards out-of-distribution generalization: A survey. *arXiv*, arXiv:2108.13624.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. In *Proceedings of the international conference on learning representations*, Banff, Canada. arXiv:1312.6199.
Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. In *Proceedings of the international conference on learning representations*, New Orleans, USA. arXiv:1805.12152.
Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2022). Open-set recognition: A good closed-set classifier is all you need? In *Proceedings of the international conference on learning representations*, Virtual. arXiv:2110.06207.
Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*, 510–520. doi:10.1038/s42256-022-00488-2
Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*, 3770. doi:10.1038/s41467-019-11786-6

Statistical prediction alone cannot identify good models of behavior

Nisheeth Srivastava, Anjali Sifar and Narayanan Srinivasan

Department of Cognitive Science, Indian Institute of Technology Kanpur, Kalyanpur, UP, India
nsrivast@iitk.ac.in
sanjali@iitk.ac.in
nsrini@iitk.ac.in
https://www.cse.iitk.ac.in/users/nsrivast/
https://sites.google.com/site/ammuns68/

doi:10.1017/S0140525X23001784, e408

Abstract

The dissociation between statistical prediction and scientific explanation advanced by Bowers et al. for studies of vision using deep neural networks is also observed in several other domains of behavior research, and is in fact unavoidable when fitting large models such as deep nets and other supervised learners, with weak theoretical commitments, to restricted samples of highly stochastic behavioral phenomena.

Bowers et al. show that, in the domain of visual perception, recent deep neural network (DNN) models that have excellent predictive performance on some types of tasks, for example, object recognition, differ from human vision in inarguable ways, for example, being biased toward making predictions based on texture rather than shape. We agree that deep-learning networks are fundamentally limited as scientific models of vision.

Generalizing Bowers et al.'s excellent observations to domains of behavior research other than vision, we suggest that throwing big models at big datasets suffers from fundamental limitations while studying scientific phenomena with low retest or inter-rater reliability (Sifar & Srivastava, 2021). In particular, large parametric models, of which supervised machine-learning models constitute an important subset, presuppose a deterministic mathematical relationship between stimuli and labels, that is, when seeing features X, the model will emit a response y. When y is stochastic, and large models are trained using one possible instance of $\{X, y\}$ observations, model predictions may actually end up becoming *too good to be true*, in the sense that they will offer statistically good predictions for phenomena that are, based on the features seen, actually unpredictable (Fudenberg, Kleinberg, Liang, & Mullainathan, 2019; Sifar & Srivastava, 2022).

Recent work has begun to quantify the notion of models being too good to be true. Fudenberg et al. (2019) define completeness as the ratio of error reduction of a model from a naive baseline to error reduction of the best possible model from the same baseline, with the best possible model defined as the table of $\{X, y\}$ mappings available in the training dataset. Since unreliable behavior intrinsically implies that the same X can correspond to more than one y, and since the model can only predict either any one of these values or an average of them, there will be some degree of irreducible error in even the best possible model.

Similarly, Sifar and Srivastava (2022) measured the retest reliability of economic preferences for risky choice using the classic "decisions from description" paradigm. They note that a basic statistical identity $\rho_{m,s2} \leq \rho_{s1,s2}\rho_{m,s1} + \sqrt{(1-\rho^2_{s1,s2})(1-\rho^2_{m,s1})}$ limits the consistency of a model m with data observed in two sessions $s1$ and $s2$. This relationship is graphically illustrated in Figure 1, showing that for low retest reliability, extremely high correlations between the model and one session's data are guaranteed to

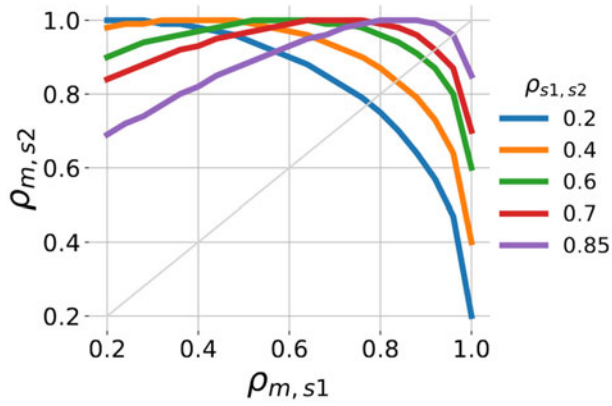


Figure 1 (Srivastava et al.). All points below the $x = y$ line on each of the curves indicate a situation where model m is guaranteed to perform worse in predicting $s2$ data when fitted to $s1$ data.

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

produce much lower correlations between that model and the other session's data, even if both sessions use the same target stimuli and protocol. Thus, the model with the best predictive accuracy when trained with one sessions' data is guaranteed to have poorer performance if tested on data collected in another session from the same participants for the same problems. Based on the measured retest reliability of economic choices, Sifar and Srivastava (2022) suggest that models showing a correlation greater than 0.85 to any given dataset may not truly be capturing important psychological phenomena about risky choices, but rather simply be overfit to dataset characteristics. Interestingly, this seems to suggest that simple generalized utility models like prospect theory are already "good enough" models of risky choice, a conclusion also reached independently by Fudenberg et al. (2019).

Notably, prediction error as observed in retest observations of a phenomenon cannot be controlled either by increasing model size or dataset size, as is prominently being recommended these days (Agrawal, Peterson & Griffiths, 2020; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021; Yarkoni & Westfall, 2017). It can only be reduced by adding more features to datasets by measuring and characterizing more sources of variability (Sifar & Srivastava, 2022). Thus, limits to predictability based on data unreliability imply that statistical model selection breaks down beyond a point for even the largest models and datasets; once multiple models can fit the data *well enough*, considerations other than goodness-of-fit must differentiate them.

In many important domains of behavior, small theory-driven models already offer predictions close to test reliability or inter-rater agreement levels in terms of accuracy (Fudenberg et al., 2019; Martin, Hofman, Sharma, Anderson, & Watts, 2016). For instance, while prospect theory is already close to an ideal model in terms of error reduction, as shown by Fudenberg et al. (2019), massive reductions in error beyond what an ideal model would be capable of are statistically claimed by large models fit to large datasets using the same impoverished feature sets that prospect theory uses (Bhatia & He, 2021; Peterson et al., 2021). The theoretical claims of such large models, however, simply offer minor modifications to the shape of the utility function used in prospect theory (Peterson et al., 2021). We argue that, in contrast to statistical predictability, scientific understanding cannot be advanced simply by fitting bigger models to bigger datasets; doing so requires fitting better models to better datasets by identifying new features that uncover additional sources of principled variation in the data.

In summary, we agree with Bowers et al. that deep-learning models, while excellent in predictive terms, may not offer unalloyedly deep insight into scientific phenomena, a trait we propose they share with other large statistical models with weak theoretical commitments endemic in many studies of behavior (Cichy & Kaiser, 2019). While the ability to search more complex function classes rather than simpler ones for models of behavior is an attractive proposition recently made possible by advances in machine learning, it is important to remain aware that using large amounts of data, with each datum generated as a restricted sample from a highly variable phenomenon, to fit highly flexible models runs the risk of obtaining high accuracy models *of the dataset* rather than the underlying *scientific phenomenon of interest*. Respecting fundamental limits to predictability of cognitive behavior must necessarily foreground mechanistic plausibility, conceptual parsimony, and consilience as criteria beyond empirical risk minimization for differentiating theoretical models.

Financial support. This work is not supported by any funding organizations.

Competing interest. None.

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(16), 8825–8835.

Bhatia, S., & He, L. (2021). Machine-generated theories of human decision-making. *Science (New York, N.Y.)*, *372*(6547), 1150–1151.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2019). Measuring the completeness of theories. *arXiv preprint* arXiv:1910.07022.

Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). Exploring limits to prediction in complex social systems. In I. Horrocks & B. Zhao (Eds.), *Proceedings of the 25th International conference on world wide web*, ACM, Montreal (pp. 683–694).

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, *372*(6547), 1209–1214.

Sifar, A., & Srivastava, N. (2021). Imprecise oracles impose limits to predictability in supervised learning. In Z. H. Zhou (Ed.), *The International joint conference on artificial intelligence (IJCAI)*, International Joint Conferences on Artificial Intelligence, Montreal (pp. 4834–4838).

Sifar, A., & Srivastava, N. (2022). Over-precise predictions cannot identify good choice models. *Computational Brain & Behavior*, *5*(3), 378–396.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Thinking beyond the ventral stream: Comment on Bowers et al.

Christopher Summerfield and Jessica A. F. Thompson

Department of Experimental Psychology, University of Oxford, Oxford, UK
christopher.summerfield@psy.ox.ac.uk
jessica.thompson@psy.ox.ac.uk
https://humaninformationprocessing.com/
https://thompsonj.github.io/about/

doi:10.1017/S0140525X23001723, e409

Abstract

Bowers et al. rightly emphasise that deep learning models often fail to capture constraints on visual perception that have been discovered by previous research. However, the solution is not to discard deep learning altogether, but to design stimuli and tasks that more closely reflect the problems that biological vision evolved to solve, such as understanding scenes and preparing skilled action.

Norbert Wiener, the founder of cybernetics, famously wrote: "the best […] model for a cat is another, or preferably the same cat" (Rosenblueth & Wiener, 1945). Wiener was referencing an assumption that good models are *general* – their predictions match data across diverse settings. A model of a cat should walk like a cat, purr like a cat, and scowl like a cat. Until recently, vision research has lacked general models. Instead, it has focused on unveiling a marvellous cabinet of perceptual curiosities, including the exotic illusions that characterise human vision. The models that explain these phenomena are typically quite narrow. For example, a model that explains *crowding* typically does not explain *filling in* and vice versa.

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

This (along with the vagaries of intellectual fashion) explains the enthusiasm that has greeted deep neural networks as theories of biological vision. Deep networks are (quite) general. After being trained to classify objects from a well-mixed distribution of natural scenes, they can generalise to accurately label new exemplars of those classes in wholly novel images. To achieve this, many networks use computational motifs recognisable from neurobiology, such as local receptivity, dimensionality reduction, divisive normalisation, and layerwise depth. This has provoked an upswell of enthusiasm around a model class that is both a passable neural simulacrum and has genuine predictive power in the natural world.

In the target article, Bowers et al. demur. Babies, they worry, may have been lost with the bathwater. Deep networks fail to capture many of the remarkable constraints on perception that have been painstakingly identified by vision researchers. The target article offers a useful tour of some behaviours we might want deep networks to display before victory is declared. For example, we should expect deep networks to show the advantage of uncrowding, to benefit from Gestalt principles, and to show a predilection to recognise objects by their shape rather than merely their texture. This point is well taken. The problem, which has been widely noted before, is that neural networks have an exasperating tendency to use every means possible to minimise their loss, including those alien to biology. In the supervised setting, if a single pixel unambiguously discloses the object label, deep networks will happily use it. If trained *ad nauseam* on shuffled labels, they will memorise the training set. If cows are always viewed in lush green pastures, they will mistake any animal in a field for a cow. None of this should be in the least surprising. It is, of course, mandated by the principles of gradient descent which empower learning in these networks.

So how do we build computational models that perceive the world in more biologically plausible ways? The target article is long on critique and short on solutions. In their concluding sections, the authors muse about the merits of a return to handcrafted models, or the augmentation of deep networks with neurosymbolic approaches. This would be a regressive step. To move away from large-scale function approximation would be to jettison the very boon that has (rightfully) propelled deep network models to prominence: Their remarkable generality.

Instead, to make progress, it would help to recall that primate vision relies on two parallel streams flowing dorsally and ventrally from early visual cortex (Mishkin, Ungerleider, & Macko, 1983). Deep networks trained for object recognition may offer a plausible model of the ventral stream, but an exclusive reliance on this stream leads to stereotyped deficits that seem to stem from a failure to understand how objects and scenes are structured. For example, damage to parieto-occipital regions can lead to integrative agnosia, where patients fail to recognise objects by integrating their parts; or to Balint's syndrome, where patients struggle to compare, count, or track multiple objects in space (Robertson, Treisman, Friedman-Hill, & Grabowecky, 1997). These are precisely the sorts of deficits that standard deep networks display: They fail to process the "objectness" of an object, relying instead on shortcuts such as mapping textures onto labels (Geirhos et al., 2020; Jagadeesh & Gardner, 2022). In primates, this computational problem is solved in the dorsal stream, where neurons code not just for objects and their labels but for the substrate (egocentric space) in which they occur. By representing space explicitly, neural populations in dorsal stream can signal how objects occupying different positions relate to each other (scene understanding), as well as encoding the spatially directed motor responses that are required to pick an object up or apprehend it with the gaze (skilled action). Thus, to account for the richness of primate visual perception, we need to build networks with both "what" and "where" streams. Recent research has started to make progress in this direction (Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021; Han & Sereno, 2022; Thompson, Sheahan, & Summerfield, 2022).

More generally, the problem is not that the deep networks are poor models of vision. The problem is that popular tests of object recognition (such as ImageNet) are unrepresentative of the challenges that biological visual systems actually evolved to solve. In the natural world, object recognition is not an end in itself, but a route to scene understanding and skilled motor control. Of course, if a network is trained to slavishly maximise its accuracy at labelling carefully curated images of singleton objects, it will find shortcuts to solving this task which do not necessarily resemble those seen in biological organisms (which generally have other more interesting things to do, such as walking, purring, and scowling).

To tackle the challenges highlighted in the target article, thus, we do not need less generality – we need more. Neuroscience researchers should focus on the complex problems that biological organisms actually face, rather than copying benchmark problems from machine learning researchers (for whom building systems that solve object recognition alone is a perfectly reasonable goal). This will require a more serious consideration of what other brain regions – including dorsal stream structures involved in spatial cognition and action selection – contribute to visual perception.

Financial support. This work was funded by the Human Brain Project (SGA3) and by a European Research Council consolidator award to C. S.

Competing interest. None.

References

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. A. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. https://papers.nips.cc/paper_files/paper/2021/file/d384dec9f5f7a64a36b5c8f03b8a6d92-Paper.pdf

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. (2020). Shortcut learning in deep neural networks. *ArXiv*. https://arxiv.org/abs/2004.07780

Han, Z., & Sereno, A. (2022). Modeling the ventral and dorsal cortical visual pathways using artificial neural networks. *Neural Computation*, *34*(1), 138–171. https://doi.org/10.1162/neco_a_01456

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(17), e2115302119. https://doi.org/10.1073/pnas.2115302119

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X

Robertson, L., Treisman, A., Friedman-Hill, S., & Grabowecky, M. (1997). The interaction of spatial and object pathways: Evidence from Balint's syndrome. *Journal of Cognitive Neuroscience*, *9*(3), 295–317. https://doi.org/10.1162/jocn.1997.9.3.295

Rosenblueth, A., & Wiener, N. (1945). The role of models in science. *Philosophy of Science*, *12*(4), 316–321. https://doi.org/10.1086/286874

Thompson, J. A. F., Sheahan, H., & Summerfield, C. (2022). Learning to count visual objects by combining "what" and "where" in recurrent memory. In *NeurIPS (gaze meets ML workshop), New Orleans*.

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

My pet pig won't fly and I want a refund

Michael J. Tarr

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
michaeltarr@cmu.edu
https://tarrlab.org

doi:10.1017/S0140525X23001760, e410

Abstract

Pigs can't fly. Any person buying a pig should understand this – it would be absurd to be upset that they can't fly or play poker. But pigs are amazing creatures and can do many interesting and useful things.

Since I retired to Florida, I have been a bit at loose ends. So, I got a pet pig. I was pretty excited. On the billboard they looked real cute, flying around on those little wings. When I picked up the piglet they told me he would get bigger fast. Sure enough, he ate like a pig. Named him PeteyPig, made a pen for him in the front yard and he rolled around in the mud and ate. But he wasn't really that cute and he sure wasn't flying. I thought, he's gotta be defective – the brochure showed pigs doing all sorts of neat things in addition to flying: carrying golf clubs, playing poker. So, I went down to the pig emporium (Fig. 1).

I said, "Hey you sold me a bad pig. He won't play poker and I am pretty sure he is never going to fly."

The salesman stared at me, then said, "Sir, you realize that those billboards and brochures are just to catch your eye? It's a pig. Pigs don't fly."

I wasn't having it. I replied, "Look here, your brochure shows poker-playing pigs, what does that mean? Petey just wallows in the mud and expects me to feed him all the time. How is that fun?"



Figure 1 (Tarr). Pigs can't fly. This image was created with the assistance of DALL·E 2 from open.ai (https://labs.openai.com).

That salesman looked worried. "Sir, you understand that is what pigs do? They don't have wings and can't play golf. That's just marketing."

Nope I thought, I am not going to be taken for a fool. "Why are you selling these pigs as pets at all? Who would want a non-flying, fat, muddy pig? You need to fix your pigs, clean them up, give em some wings, and teach them to play poker!"

The salesman gave me another look. "Look sir, pigs are great. They can do amazing things. A miracle of nature. They're playful and smarter than dogs. But pigs will do what pigs will do. Complain all you want, but they won't fly. If you want a golf club-toting poker buddy, hire someone. Flying is out."

Well, I had heard enough. I walked straight back to the Villages. Petey was in the front yard, covered in mud. I tossed him some carrots and sat down. Ever hopeful, I said, "Petey old buddy, let me show you the queen of hearts…."

Bowers et al. build a straw house by motivating their arguments through quotes that are more marketing than scientific claims. Much like our protagonist, we need to be smart consumers of science. I don't think there is much actual confusion that deep neural networks (DNNs) are "models of the human visual system." Rather, like the computer vision models that preceded DNNs, they serve as "proxy models" that surface the role(s) of assumptions and constraints in complex systems (Leeds, Seibert, Pyles, & Tarr, 2013).

As proxy models, DNNs are remarkable because of what they can do in comparison with prior models. DNNs learn task-relevant representations that are often well aligned with representations in neural systems that support a common task (Yamins & DiCarlo, 2016). This level of alignment is a dramatic shift from the mostly much poorer attempts to account for neural data that preceded DNNs. Even so, it should be obvious that DNNs, in and of themselves, don't have many of the characteristics that define intelligence in biological systems.

As a field we should have a productive discussion about what inferences we can draw from DNNs and other computational models (Guest & Martin, 2023). However, such discussions should involve less hyperbole ("Deep problems…") and less handwringing about what current models can't do; instead, they should focus on what DNNs can do. They might be pigs, they will never fly, but they can do some pretty cool stuff. We should figure out how and why.

Acknowledgments. The author thanks Marlene Behrmann, David Plaut, and Rob Kass for their comments.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sector.

Competing interest. None.

References

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*, 213–227. doi:10.1007/s42113-022-00166-x

Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, *13*(13), 25. doi:10.1167/13.13.25

Yamins, D., & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365. doi:10.1038/nn.4244

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

Neural networks, AI, and the goals of modeling

Walter Veit[a] and Heather Browning[b]

[a]Department of Philosophy, University of Bristol, Bristol, UK and [b]Department of Philosophy, University of Southampton, Southampton, UK
wrwveit@gmail.com
DrHeatherBrowning@gmail.com
https://walterveit.com/
https://www.heatherbrowning.net/

doi:10.1017/S0140525X23001681, e411

Abstract

Deep neural networks (DNNs) have found many useful applications in recent years. Of particular interest have been those instances where their successes imitate human cognition and many consider artificial intelligences to offer a lens for understanding human intelligence. Here, we criticize the underlying conflation between the predictive and explanatory power of DNNs by examining the goals of modeling.

As is often the case with technological and computational progress, our newest and most sophisticated tools come to be seen as models for human cognition. What perhaps began with Gottfried Leibniz – who famously compared the mind to a mill – has a long philosophical, and now cognitivist, tradition. While it is natural to draw inspiration from technological progress to advance our understanding of the mind, unsurprisingly there are many staunch critics of the idea that the human mind should be seen as anything like a computer, with only a difference in substance. In their target article, Bowers et al. offer a compelling instance of this general criticism, arguing against recent attempts to describe deep neural networks (DNNs) as the best models for understanding human vision (or any form of biological vision).

While DNNs have admittedly been extremely successful at classifying objects on the basis of photographs – indeed even exceeding human levels of performance in some domains – Bowers et al. essentially argue that they have very little explanatory power for human vision, due to having little in common with the mechanisms of biological vision. In order to improve our understanding of human vision, they instead advocate focusing more on explaining actual psychological findings by offering testable hypotheses.

This argument is reminiscent of many other scientific debates, such as whether artificial neural networks constitute a good model for the human brain more generally (Saxe, Nelli, & Summerfield, 2021; Schaeffer, Khona, & Fiete, 2022). It also has links to long-standing discussions in the philosophy of science on the goals of science, between those that seek successful predictions and those that seek out true explanations – a debate that is sometimes framed as instrumentalists versus realists (see Psillos, 2005). While scientists may not frame their disagreement in exactly these terms, their arguments may similarly be reflective of very different attitudes toward the methodology and theoretical assumptions of their disciplines.

Our goal here is not to argue against the view provided by Bowers et al. Indeed, we strongly agree with their general argument that the predictive power of DNNs is insufficient to vindicate their status as models for biological vision. Even highly theoretical work has to make contact with empirical findings to promote greater explanatory power of the models. Instead, our aim here will be to take a philosophy of science perspective to examine the goals of modeling, illuminating where the disagreements between scientists in this area originate.

First, there is the concern of conflating prediction with explanation. While some early philosophers of science maintained that prediction and explanation are formally (almost) equivalent, this view was quickly challenged (Rescher, 1958) and today is almost universally rejected within philosophy of science. Nevertheless, in many scientific disciplines there is still a continuous and common conflation between the predictive power of a model and its explanatory power. Thus, we should not be surprised at all that many scientists have made the jump from the striking predictive success of DNNs to the bolder claim that they are representative models of human vision. While predictive power can certainly constitute one piece of good evidence for one model having greater explanatory power than another, this relationship is not guaranteed. This is especially the case when we make extrapolations from machine learning to claims about the mechanisms behind how biological agents learn and categorize the world. As Bowers et al. point out, the current evidence does not support such a generalization and instead suggests there are more likely to be dissimilar causal mechanisms underlying the observed patterns.

Second, as philosophers of biology have argued for the last several decades, many of the properties and abilities of biological systems can be multiply realized, that is, they can be realized through different causal mechanisms (Ross, 2020; Sober, 1999). Thus, the idealizations within one model may not be adequate for its application in a different target system. Just because DNNs are the first artificial intelligences (AIs) we have created that approximate human levels of success in vision (or cognition) does not mean that biological systems must be operating under the same principles. Indeed, the different origins and constraints on developing DNNs as compared with the evolution of human vision mean that this is even less likely to be the case.

Third, the authors' emphasis on controlled experiments that help us to understand mechanisms by manipulating independent variables is an important one and one that has been a common theme in recent work in the philosophy of science (e.g., Schickore, 2019). This is a very different enterprise than the search for the best predictive models and AI researchers will benefit greatly from taking note of this literature. Part of the hype about AI systems has precisely been due to the confusion between predictive power and explanatory causal understanding. Prediction can be achieved through a variety of means, many of which will not be sufficiently relevantly similar to provide a good explanation.

We wish to finish by pointing out that the inadequacy of DNNs for understanding biological vision is not at all an indictment of their usefulness for other purposes. Science operates under a plurality of models and these will inevitably have different goals (Veit, 2019). It is particularly interesting that DNNs have outperformed humans in some categorization tasks, since it suggests that artificial neural networks do not have to operate in the same ways as biological vision in order to imitate or even trump its successes. Indeed, there is still an important explanatory question to answer: If DNNs could constitute a superior form of visual processing, why have biological systems evolved different ways of categorizing the world? To answer these and related questions,

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

scientists will have to seek greater collaboration and integration with psychological and neurological research, as suggested by Bowers et al. As we thus hope to have made clear here, this debate would greatly benefit by further examining its underlying methodological and philosophical assumptions as well as engaging with the literature in philosophy of science where these issues have been discussed at length.

Financial support. This study is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 101018533).

Competing interest. None.

References

Psillos, S. (2005). *Scientific realism: How science tracks truth*. Routledge.
Rescher, N. (1958). On prediction and explanation. *The British Journal for the Philosophy of Science*, *8*(32), 281–290.
Ross, L. N. (2020). Multiple realizability from a causal perspective. *Philosophy of Science*, *87*(4), 640–662.
Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67.
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*, 2022-08.
Schickore, J. (2019). The structure and function of experimental control in the life sciences. *Philosophy of Science*, *86*(2), 203–218.
Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of Science*, *66*(4), 542–564.
Veit, W. (2019). Model pluralism. *Philosophy of the Social Sciences*, *50*(2), 91–114.

Neither hype nor gloom do DNNs justice

Felix A. Wichmann[a], Simon Kornblith[b] and Robert Geirhos[b]

[a]Neural Information Processing Group, University of Tübingen, Tübingen, Germany and [b]Google Research, Brain Team, Toronto, ON, Canada
felix.wichmann@tuebingen.de
skornblith@google.com
geirhos@google.com

doi:10.1017/S0140525X23001711, e412

Abstract

Neither the hype exemplified in some exaggerated claims about deep neural networks (DNNs), nor the gloom expressed by Bowers et al. do DNNs as models in vision science justice: DNNs rapidly evolve, and today's limitations are often tomorrow's successes. In addition, providing explanations as well as prediction and image-computability are model desiderata; one should not be favoured at the expense of the other.

We agree with Bowers et al. that some of the quoted statements at the beginning of their target article about deep neural networks (DNNs) as "best models" are exaggerated – perhaps some of them bordering on scientific hype (Intemann, 2020). However, only the authors of such exaggerated statements are to blame, not DNNs: Instead of blaming DNNs, perhaps Bowers et al. should have engaged in a critical discussion of the increasingly widespread practice of rewarding impact and boldness over carefulness and modesty that allows hyperbole to flourish in science. This is unfortunate as the target article does mention a number of valid issues with DNNs in vision science and raises a number of valid concerns. For example, we fully agree that human vision is much more than recognising photographs of objects in scenes; we also fully agree there are still a number of important behavioural differences between DNNs and humans even in terms of core object recognition (DiCarlo, Zoccolan, & Rust, 2012), that is, even when recognising photographs of objects in scenes, such as DNNs' adversarial susceptibility (target article, sect. 4.1.1) or reliance on local rather than global features (target article, sect. 4.1.3). However, we do not subscribe to the somewhat gloomy view of DNNs in vision science expressed by Bowers et al. We believe that image-computable models are essential to the future of vision science, and DNNs are currently the most promising – albeit not yet fully adequate – model class for core object recognition.

Importantly, any behavioural differences between DNNs and humans can only be a snapshot in time – true as of today. Unlike Bowers et al. we do not see any evidence that future, novel DNN architectures, training data and regimes may not be able to overcome at least some of the limitations mentioned in the target article – and Bowers et al. certainly do not provide any convincing evidence why solving such tasks is beyond DNNs *in principle, that is, forever*. In just over a decade, DNNs have come a long way from AlexNet, and we still witness tremendous progress in deep learning. Until recently, DNNs lacked robustness to image distortions; now some match or outperform humans on many of them. DNNs made very different error patterns than humans; newer models achieve at least somewhat better consistency (Geirhos et al., 2021). DNNs used to be texture-biased; now some are shape-biased similar to humans (Dehghani et al., 2023). With DNNs, today's limitations are often tomorrow's success stories.

Yes, current DNNs still fail on a large number of "psychological tasks," from (un-)crowding (Doerig, Bornet, Choung, & Herzog, 2020) to focusing on local rather than global shape (Baker, Lu, Erlikhman, & Kellman, 2018), from similarity judgements (German & Jacobs, 2020) to combinatorial judgements (Montero, Bowers, Costa, Ludwig, & Malhotra, 2022); furthermore, current DNNs lack (proper, human-like) sensitivity to Gestalt principles (Biscione & Bowers, 2023). But current DNNs in vision are typically trained to recognise static images; their failure on "psychological tasks" without (perhaps radically) different training or different optimisation objectives does not surprise us – just as we do not expect a traditional vision model of motion processing to predict lightness induction or an early spatial vision model to predict Gestalt laws, at least not without substantial modification and fitting it to suitable data. To overcome current DNNs' limitations on psychological tasks we need more DNN research inspired by vision science, not just engineering to improve models' overall accuracy – here we certainly agree again with Bowers et al.

Moreover, for many of the abovementioned psychological tasks, there simply do not exist successful traditional vision models. Why single out DNNs as failures if no successful computational model exists, at least not image-computable models? Traditional "object recognition" models only model isolated aspects of object recognition, and it is difficult to tell how well they model these aspects, since only image-computable models can actually recognise objects. Here, image-computability is far

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

more than just a "nice to have" criterion since it facilitates falsifiability. We think that Bowers et al.'s long list of DNN failures should rather be taken as a list of desiderata of what future image-computable models of human vision should explain and predict.

Although we do not know whether DNNs will be sufficient to meet this challenge, only future research will resolve the many open questions: Is our current approach of applying predominantly discriminative DNNs as computational models of human vision sufficient to obtain truly successful models? Do we need to incorporate, for example, causality (Pearl, 2009), or generative models such as predictive coding (Rao & Ballard, 1999) or even symbolic computations (Mao, Gan, Kohli, Tenenbaum, & Wu, 2019)? Do we need to ground learning in intuitive theories of physics and psychology (Lake, Ullman, Tenenbaum, & Gershman, 2017)?

Finally, it appears as if Bowers et al. argue that models should first and foremost provide explanations, as if predictivity – which includes but is not limited to image-computability – did not matter much. (Or observational data; successful models need to be able to explain and predict data from hypothesis-driven experiments as well as observational data.) While we agree with Bowers et al. that in machine learning there is a tendency to blindly chase improved benchmark numbers without seeking understanding of underlying phenomena, we believe that both prediction and explanation are required: An explanation without prediction cannot be trusted, and a prediction without explanation does not aid understanding. What we need is not a myopic focus on one or the other, but to be more explicit about modelling goals – both in the target article by Bowers et al. and in general, as we argue in a forthcoming article (Wichmann & Geirhos, 2023).

We think that neither the hype exemplified in some exaggerated claims about DNNs, nor the gloom expressed by Bowers et al. do DNNs and their application to vision science justice. Looking forward, if we want to make progress towards modelling and understanding human visual perception, we believe that it will be key to move beyond both hype and gloom and carefully explore similarities and differences between human vision and rapidly evolving DNNs.

Financial support. This work was supported by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 4, project number: 276693517 to F. A. W. In addition, F. A. W. is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

Competing interest. None.

References

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.

Biscione, V., & Bowers, J. S. (2023). Mixed evidence for Gestalt grouping in deep neural networks. *Computational Brain & Behavior*, 1–19. https://doi.org/10.1007/s42113-023-00169-2

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

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research*, *167*, 39–45.

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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 23885–23899). Curran Associates.

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

Intemann, K. (2020). Understanding the problem of "hype": Exaggeration, values, and trust in science. *Canadian Journal of Philosophy*, *52*(3), 1–16.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations* (ICLR), New Orleans, Louisiana, United States, pp. 1–28. https://iclr.cc/Conferences/2019

Montero, M. L., Bowers, J. S., Costa, R. P., Ludwig, C. J. H., & Malhotra, G. (2022). Lost in latent space: Disentangled models and the challenge of combinatorial generalisation. *arXiv*, arXiv:2204.02283.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioural models of human visual perception? *Annual Review of Vision Science*, *9*. https://doi.org/10.1146/annurev-vision-120522-031739

Using DNNs to understand the primate vision: A shortcut or a distraction?

Yaoda Xu[a] and Maryam Vaziri-Pashkam[b]

[a]Department of Psychology, Yale University, New Haven, CT, USA and [b]National Institute of Mental Health, Bethesda, MD, USA
yaoda.xu@yale.edu, https://sites.google.com/view/yaodaxu/home
maryam.vaziri-pashkam@nih.gov, https://mvaziri.github.io/Homepage/Bio.html

doi:10.1017/S0140525X23001528, e413

Abstract

Bowers et al. bring forward critical issues in the current use of deep neural networks (DNNs) to model primate vision. Our own research further reveals fundamentally different algorithms utilized by DNNs for visual processing compared to the brain. It is time to reemphasize the value of basic vision research and put more resources and effort on understanding the primate brain itself.

Similarities exist between deep neural networks (DNNs) and the primate brain in how they process visual information. This has generated the excitement that perhaps the algorithms governing high-level vision would "automagically" emerge in DNNs to provide us with a shortcut to understand and model primate vision. In their detailed critiques, Bowers et al. bring forward significant drawbacks in the current applications of DNNs to explain primate vision. Perhaps it is time to step back and ask: Is it really a shortcut or a distraction to use DNNs to understand the primate vision?

Using detailed examples, Bowers et al. pointed out that performance alone does not constitute as good evidence that the same processing algorithms are utilized by both the primate brain and DNNs. They showed that DNNs fail to account for

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

a large number of findings in vision research. In our own research, by comparing DNN responses to our previously collected fMRI datasets (Vaziri-Pashkam, Taylor, & Xu, 2019; Vaziri-Pashkam & Xu, 2019), we found that DNNs' performance is related to the fact that they are built following the known architecture of the primate lower visual areas and are trained with real-world object images. Consequently, DNNs are successful at fully capturing the visual representational structures of lower human visual areas during the processing of real-world images, but not those of higher human visual areas during the processing of these images or that of artificial images at either level of processing (Xu & Vaziri-Pashkam, 2021a). The close brain–DNN correspondence found in earlier fMRI studies appears to be overly optimistic by including only real-world objects and compared to brain data with relatively lower power. When we expanded the comparisons to a broader set of real-world stimuli and to artificial stimuli as well as comparing to brain data with a higher power, this correspondence becomes much weaker.

Perhaps the most troubling finding from our research is that DNNs do not form the same transformation-tolerant visual object representations as the human brain does. Decades of neuroscience research has shown that one of the greatest achievements of primate high-level vision is its ability to extract object identity among changes in nonidentity features to form transformation-tolerant object representations (DiCarlo & Cox, 2007; DiCarlo, Zoccolan, & Rust, 2012; Tacchetti, Isik, & Poggio, 2018). This allows us to rapidly recognize an object under different viewing conditions. Computationally, achieving tolerance reduces the complexity of learning by requiring fewer training examples and improves generalization to objects and categories not included in training (Tacchetti et al., 2018). We found that while the object representational geometry was increasingly tolerant to changes in nonidentity features from lower to higher human visual areas, this was not the case in DNNs pretrained for object classification regardless of network architecture, depth, with/without recurrent processing, or with/without pretraining to emphasize shape processing (Xu & Vaziri-Pashkam, 2022). By comparing DNN responses with another existing fMRI dataset (Jeong & Xu, 2017), we further showed that while the human higher visual areas exhibit clutter tolerance, such that fMRI responses to an object pair can be well approximated by the average responses to each constituent object shown alone, this was not the case in DNNs (Mocz, Jeong, Chun, & Xu, 2023). We additionally found that DNNs differ from the human visual areas in how they represent object identity and nonidentity features over the course of visual processing (Xu & Vaziri-Pashkam, 2021b). With their vast computing power, DNNs likely associate different instances of an object to a label without preserving the object representational geometry across nonidentity feature changes to form brain-like tolerance. While this is one way to achieve tolerance, it requires a large number of training data and has a limited ability to generalize to objects not included in the training, the two major drawbacks associated with the current DNNs (Serre, 2019).

If DNNs use fundamentally different algorithms for visual processing, then in what way do they provide shortcuts, rather than distractions, to help us understand primate vision? It may be argued that since DNNs are the current best models in producing human-like behavior, we should keep refining them using our knowledge of the primate brain. This practice, however, relies on a thorough understanding of the primate brain. If we could already accomplish this, do we still need DNN modeling? As stated by Kay (2018), given that DNNs typically contain millions or even hundreds of millions of free parameters, even if we are successful in duplicating the primate brain in DNNs, how does replacing one black box (the primate brain) with another black box (a DNN) constitute a fundamental understanding of primate vision? Perhaps it is time to reemphasize the value of basic vision and neuroscience research and put more effort and resources on understanding the precise algorithms used by the primate brain in visual processing.

While current DNNs may not provide an easy and quick shortcut to understanding primate vision, can they still be useful? Some have used DNNs to test our theories about the topographic (Blauch, Behrmann, & Plaut, 2022) and anatomical organization of the brain (Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021) and to answer "why" brains work the way they do (Kanwisher et al., 2023). Here again, when our theories about the brain are not borne out in DNNs, are our theories wrong or are DNNs just ill models in those regards? It remains to be seen if such approaches can bring us fundamental understanding of the brain beyond what we already know. Although DNNs may yet to possess the explanation power we desire, they could nevertheless serve as powerful simulation tools to aid vision research. For example, we have recently used DNNs to fine tune our visual stimuli and help us lay out the detailed analysis pipeline that we plan to use to study visual processing in the human brain (e.g., Tang, Chin, Chun, & Xu, 2022; Taylor & Xu, 2021). DNNs are likely here to stay. Understanding their drawbacks and finding the right way to harness their power will be the key for future vision research.

Author's contribution. Y. X. wrote the manuscript with comments from M. V.-P.

Financial support. Y. X. was supported by the National Institute of Health (NIH) Grant 1R01EY030854. M. V.-P. was supported by NIH Intramural Research Program ZIA MH002035.

Competing interest. None.

References

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *119*, e2112566119.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Science*, *11*, 333–341.

DiCarlo, J. J., Zoccolan, D., & Rust, R. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*, 415–434.

Jeong, S. K., & Xu, Y. (2017). Task-context dependent linear representation of multiple visual objects in human parietal cortex. *Journal of Cognitive Neuroscience*, *29*, 1778–1789.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neuroscience*, *46*, 240–254.

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*, 101–109.

Mocz, V., Jeong, S. K., Chun, M., & Xu, Y. (2023). The representation of multiple visual objects in human ventral visual areas and in convolutional neural networks. *Scientific Reports*, *13*, 9088.

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, *5*, 399–426.

Tacchetti, A., Isik, L., & Poggio, T. A. (2018). Invariant recognition shapes neural representations of visual input. *Annual Review of Vision Science*, *4*, 403–422.

Tang, K., Chin, M., Chun, M., & Xu, Y. (2022). The contribution of object identity and configuration to scene representation in convolutional neural networks. *PLoS ONE*, *17*, e0270667.

Taylor, J., & Xu, Y. (2021). Joint representation of color and shape in convolutional neural networks: A stimulus-rich network perspective. *PLoS ONE*, *16*, e0253442.

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

Vaziri-Pashkam, M., Taylor, J., & Xu, Y. (2019). Spatial frequency tolerant visual object representations in the human ventral and dorsal visual processing pathways. *Journal of Cognitive Neuroscience*, *31*, 49–63.
Vaziri-Pashkam, M., & Xu, Y. (2019). An information-driven 2-pathway characterization of occipitotemporal and posterior parietal visual object representations. *Cerebral Cortex*, *29*, 2034–2050.
Xu,Y.,&Vaziri-Pashkam,M.(2021a).Limitedcorrespondenceinvisualrepresentationbetween the human brain and convolutional neural networks. *Nature Communications*, *12*, 2065.
Xu, Y., & Vaziri-Pashkam, M. (2021b). The coding of object identity and nonidentity features in human occipito-temporal cortex and convolutional neural networks. *Journal of Neuroscience*, *41*, 4234–4252.
Xu, Y., & Vaziri-Pashkam, M. (2022). Understanding transformation tolerant visual object representations in the human brain and convolutional neural networks. *NeuroImage*, *263*, 119635.

Why psychologists should embrace rather than abandon DNNs

Galit Yovel[a,b] and Naphtali Abudarham[a]

[a]School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel and [b]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
gality@tauex.tau.ac.il; https://people.socsci.tau.ac.il/mu/galityovel/
naphtool@gmail.com

doi:10.1017/S0140525X2300167X, e414

Abstract

Deep neural networks (DNNs) are powerful computational models, which generate complex, high-level representations that were missing in previous models of human cognition. By studying these high-level representations, psychologists can now gain new insights into the nature and origin of human high-level vision, which was not possible with traditional handcrafted models. Abandoning DNNs would be a huge oversight for psychological sciences.

Computational modeling has long been used by psychologists to test hypotheses about human cognition and behavior. Prior to the recent rise of deep neural networks (DNNs), most computational models were handcrafted by scientists who determined their parameters and features. In vision sciences, these models were used to test hypotheses about the mechanisms that enable human object recognition. However, these handcrafted models used simple, engineered-designed features (e.g., Gabor wavelets), which produced low-level representations that did not account for human-level, view-invariant object recognition (Biederman & Kalocsai, 1997; Turk & Pentland, 1991). The main advantage of DNNs over these traditional models is not only that they reach human-level performance in object recognition, but that they do so through hierarchical processing of the visual input that generates high-level, view-invariant visual features. These high-level features are the "*missing link*" between the low-level representations of the hand crafted models and human-level object classification. They therefore offer psychologists an unprecedented opportunity to test hypotheses about the origin and nature of these high-level representations, which were not available for exploration so far.

In the target article, Bowers et al. propose that psychologists should abandon DNNs as models of human vision, because they do not produce some of the perceptual effects that are found in humans. However, many of the listed perceptual effects that DNNs fail to produce are also not produced by the traditional handcrafted computational vision models, which have been prevalently used to model human vision. Furthermore, although current DNNs are primarily developed for engineering purposes (i.e., best performance), there are myriad of ways in which they could and should be modified to better resemble the human mind. For example, current DNNs that are often used to model human face and object recognition (Khaligh-Razavi & Kriegeskorte, 2014; O'Toole & Castillo, 2021; Yamins & DiCarlo, 2016) are trained on static images (Cao, Shen, Xie, Parkhi, & Zisserman, 2018; Deng et al., 2009), whereas human face and object recognition are performed on continuous streaming of dynamic, multi-modal information. One way that has recently been suggested to close this gap is to train DNNs on movies that are generated by head-mounted cameras attached to infants' forehead (Fausey, Jayaraman, & Smith, 2016), to better model the development of human visual systems (Smith & Slone, 2017). Training DNNs initially on blurred images also provided insights into the potential advantage of the initial low acuity of infants' vision (Vogelsang et al., 2018). Such and many other modifications (e.g., multi-modal self-supervised image-language training, Radford et al., 2021) in the way DNNs are built and trained may generate perceptual effects that are more human-like (Shoham, Grosbard, Patashnik, Cohen-Or, & Yovel, 2022). Yet even current DNNs can advance our understanding of the nature of the high-level representations that are required for face and object recognition (Abudarham, Grosbard, & Yovel, 2021; Hill et al., 2019), which are still undefined in current neural and cognitive models. This significant computational achievement should not be dismissed.

Bowers et al. further claim that DNNs should be used to test hypotheses rather than to solely make predictions. We fully agree and further propose that psychologists are best suited to apply this approach by utilizing the same procedures they have used for decades to test hypotheses about the hidden representations of the human mind. Since the early days of psychological sciences, psychologists have developed a range of elegant experimental and stimulus manipulations to study human vision. The same procedures can now be used to explore the nature of DNNs' high-level hidden representations as potential models of the human mind (Ma & Peters, 2020). For example, the *face inversion effect* is a robust, extensively studied, and well-established effect in human vision, which refers to the disproportionally large drop in performance that humans show for upside-down compared to upright faces (Cashon & Holt, 2015; Farah, Tanaka, & Drain, 1995; Yin, 1969). Because the low-level features extracted by, handcrafted algorithms are similar for upright and inverted faces, these traditional models do not reproduce this effect. Interestingly, a human-like face inversion effect that is larger than an object inversion effect is found in DNNs (Dobs, Martinez, Yuhan, & Kanwisher, 2022; Jacob, Pramod, Katti, & Arun, 2021; Tian, Xie, Song, Hu, & Liu, 2022; Yovel, Grosbard, & Abudarham, 2023). Thus, we can now use the same stimulus and task manipulations that were used to study this effect in numerous human studies, to test hypotheses about the mechanism that may underlie this perceptual effect. Moreover, by manipulating DNNs' training diet, we can examine what type of experience is needed to generate this human-like perceptual effect, which is impossible to test in humans where we have no control over their perceptual experience. Such an approach has recently been used to address a long-lasting debate in cognitive sciences about the domain-specific versus the expertise hypothesis in face recognition (Kanwisher, Gupta, & Dobs, 2023; Yovel et al., 2023).

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

It was psychologists, not engineers, who first designed these neural networks to model human intelligence (McClelland, McNaughton, & O'Reilly, 1995; Rosenblatt, 1958; Rumelhart, Hinton, & Williams, 1986). It took more than 60 years since the psychologist, Frank Rosenblatt published his report about the *perceptron*, for technology to reach its present state where these hierarchically structured algorithms can be used to study the complexity of human vision. Abandoning DNNs would be a huge oversight for cognitive scientists, who can contribute considerably to the development of more human-like DNNs. It is therefore pertinent that psychologists join the artificial intelligence (AI) research community and study these models in collaboration with engineers and computer scientists. This is a unique time in the history of cognitive sciences, where scientists from these different disciplines have shared interests in the same type of computational models that can advance our understanding of human cognition. This opportunity should not be missed by psychological sciences.

Financial support. This study was funded by an ISF 971/21 and Joint NSFC-ISF 2383/18 to G. Y.

Competing interest. None.

References

Abudarham, N., Grosbard, I., & Yovel, G. (2021). Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *Cognitive Science*, *45*(9), e13031. https://doi.org/10.1111/cogs

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1358), 1203–1219. https://doi.org/10.1098/rstb.1997.0103

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE international conference on automatic face and gesture recognition*, FG 2018 (pp. 67–74). https://doi.org/10.1109/FG.2018.00020

Cashon, C. H., & Holt, N. A. (2015). Developmental origins of the face inversion effect. In *Advances in child development and behavior* (1st ed., Vol. 48, pp. 117–150). Elsevier. https://doi.org/10.1016/bs.acdb.2014.11.008

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database (pp. 248–255). https://doi.org/10.1109/cvprw.2009.5206848

Dobs, K., Martinez, J., Yuhan, K., & Kanwisher, N. (2022). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, *120*(32), e2220642120.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 628–634. https://doi.org/10.1037/0096-1523.21.3.628

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107. https://doi.org/10.1016/j.cognition.2016.03.005

Hill, M. Q., Parde, C. J., Castillo, C. D., Colón, Y. I., Ranjan, R., Chen, J.-C., … O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, *1*(11), 522–529. https://doi.org/10.1038/s42256-019-0111-7

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. https://doi.org/10.1038/s41467-021-22078-3

Kanwisher, N., Gupta, P., & Dobs, K. (2023). CNNs reveal the computational implausibility of the expertise hypothesis. *iScience*, *26*(2), 105976. https://doi.org/10.1016/j.isci.2023.105976

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

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: Towards using deep nets as models for human behavior (pp. 1–39).

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

O'Toole, A. J., & Castillo, C. D. (2021). Face recognition by humans and machines: Three fundamental advances from deep learning. *Annual Review of Vision Science*, *7*, 543–570. https://doi.org/10.1146/annurev-vision-093019-111701

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Shoham, A., Grosbard, I., Patashnik, O., Cohen-Or, D., & Yovel, G. (2022). Deep learning algorithms reveal a new visual-semantic representation of familiar faces in human perception and memory. *Biorxiv*, 2022-10.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, *8*, 1–10. https://doi.org/10.3389/fpsyg.2017.02124

Tian, F., Xie, H., Song, Y., Hu, S., & Liu, J. (2022). The face inversion effect in deep convolutional neural networks. *Frontiers in Computational Neuroscience*, *16*, 1–8. https://doi.org/10.3389/fncom.2022.854218

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44), 11333–11338. https://doi.org/10.1073/pnas.1800901115

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141.

Yovel, G., Grosbard, I., & Abudarham, N. (2023). Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proceedings of the Royal Society B*, *290*(1998), 20230093.

Authors' Response

Clarifying status of DNNs as models of human vision

Jeffrey S. Bowers[a], Gaurav Malhotra[a], Marin Dujmović[a], Milton L. Montero[a], Christian Tsvetkov[a], Valerio Biscione[a], Guillermo Puebla[b], Federico Adolfi[c], John E. Hummel[d], Rachel F. Heaton[d], Benjamin D. Evans[e], Jeffrey Mitchell[e] and Ryan Blything[f]

[a]School of Psychological Science, University of Bristol, Bristol, UK; [b]National Center for Artificial Intelligence, Macul, Chile; [c]Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt am Main, Germany; [d]Psychology Department, University of Illinois Urbana–Champaign, Champaign, IL, USA; [e]Department of Informatics, School of Engineering and Informatics, University of Sussex, Brighton, UK and [f]School of Psychology, Aston University, Birmingham, UK

j.bowers@bristol.ac.uk; https://jeffbowers.blogs.bristol.ac.uk/
gaurav.malhotra@bristol.ac.uk
marin.dujmovic@bristol.ac.uk
m.lleramontero@bristol.ac.uk
christian.tsvetkov@bristol.ac.uk
valerio.biscione@gmail.com
guillermo.puebla@bristol.ac.uk
fedeadolfi@gmail.com
jehummel@illinois.edu
rmflood2@illinois.edu
b.d.evans@sussex.ac.uk
j.mitchell@napier.ac.uk
r.blything@aston.ac.uk

doi:10.1017/S0140525X23002777, e415

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

Abstract

On several key issues we agree with the commentators. Perhaps most importantly, everyone seems to agree that psychology has an important role to play in building better models of human vision, and (most) everyone agrees (including us) that deep neural networks (DNNs) will play an important role in modelling human vision going forward. But there are also disagreements about what models are for, how DNN–human correspondences should be evaluated, the value of alternative modelling approaches, and impact of marketing hype in the literature. In our view, these latter issues are contributing to many unjustified claims regarding DNN–human correspondences in vision and other domains of cognition. We explore all these issues in this response.

R1. Overview

We are pleased that so many commentators agree with so many of our core claims. For instance, there is general agreement that current deep neural networks (DNNs) do a poor job in accounting for many psychological findings; that an important direction for future research is to train DNNs on new tasks and datasets that more closely capture human experience; and that new objective functions like self-supervision may improve DNN–human correspondences. Most importantly, there is widespread agreement that research in psychology should play a central role in building better models of human vision. It is important to appreciate the implication of this last point because psychological experiments reveal some weird and wonderful properties of human vision that DNNs must seek to explain. We start by discussing some of these key properties before responding to the specific points of the commentators.

To give only the most cursory of overviews, the following findings should play a central role in theory and model building. The input to our visual system is degraded due to a large blind spot and an inverted retina with light having to pass through multiple layers of retinal neurons, axons, and blood vessels before reaching the photoreceptors. Nevertheless, we are unaware of the degraded signals due to a process of actively filling in missing signals in early visual cortex (e.g., Grossberg, 2003; Ramachandran & Gregory, 1991). We have fovea that supports high-acuity colour vision for only about 2 degrees of visual angle (about the size of a thumbnail at arm's length). Nevertheless, we have the subjective sense of a rich visual experience across a much wider visual field because we move our eyes approximately three times per second (Rayner, 1978), with the encoding of visual inputs suppressed during each saccade (Matin, 1974), and the visual system somehow integrating inputs across fixations (Irwin, 1991). At the same time, we can identify multiple objects in scenes following a single fixation (Biederman, 1972), with object identification taking approximately 150 ms (Thorpe, Fize, & Marlot, 1996) – too quick to rely on recurrence. We are also blind to major changes in a scene as revealed by change blindness (Simons & Levin, 1997) and have a visual short-term memory of approximately four items (Cowan, 2001). Our visual system organizes image contours by various Gestalt rules to separate figure from ground (Wagemans et al., 2012) and organize contours to build representations of object parts (Biederman, 1987). Objects are encoded in terms of their surfaces, parts, and relations between parts to build three-dimensional (3D) representations relying on monocular and binocular inputs (Biederman, 1987; Marr, 1982; Nakayama & Shimojo, 1992). Colour, form, and motion processing are factorized to the extent that it is possible to be cortically colour blind (Cavanagh et al., 1998), or suffer motion blindness where objects disappear during motion but are visible and recognizable while static (Zeki, 1991), or show severe impairments with object identification while maintaining the ability to reach and manipulate objects (Goodale & Milner, 1992). Participants can even classify objects while denying seeing them (**Koculak & Wierzchoń**). Our visual system manifests a wide range of visual, size, and shape constancies to estimate the distal properties of the world independent of the lighting and object pose, and we suffer from size, colour, and motion illusions that reflect the very mechanisms that serve the building of these distal representations from the proximal image projected onto our retinas. These representations of distal stimuli in the world support a range of visual tasks, including object classification, navigation, grasping, and visual reasoning. All this is done with spiking networks composed of neurons with a vast range of morphologies that vary in ways relevant to their function, with architectures constrained by evolution and biophysics.

All of this and much more needs to be explained, and various modelling approaches are warranted. We agree with the commentators that one valuable approach is to keep working with current image-computable DNNs while altering the tasks they solve, the data they are fed, their objective functions, learning rules, and architectures. Perhaps DNNs will converge with the biological solutions in some important respects. Whether DNNs will "automagically" (**Xu & Vaziri-Pashkam**) converge on many of these solutions when trained on the right tasks and data, however, is far from certain, and in our view, it is a mistake to put all our eggs in this one basket. Whatever approach one adopts, the current trend of emphasizing prediction success on observational behavioural and brain benchmarks and downplaying failures is unlikely to advance our understanding of human vision and the brain more generally.

Our response to the commentaries is organized as follows. In section R2 we show there is no basis for the claim that we are advocating for the abandonment of DNNs as a modelling framework to test hypotheses about human vision. In sections R3 and R4 we challenge the common claim that image computability is the minimal criteria for any serious model of vision and that DNNs are the "current best" models of human vision. In section R5 we argue that models should be developed for the sake of explanations rather than predictions. In section R6 we discuss how the marketing of DNNs as the best models of human vision is contributing to a current trend of emphasizing DNN–human similarities and downplaying discrepancies. Finally, in section R7, we respond to the **DiCarlo, Yamins, Ferguson, Fedorenko, Bethge, Bonnen, & Schrimpf (DiCarlo et al.)** and **Golan, Taylor, Schütt, Peters, Sommers, Seeliger, Doerig, Linton, Konkle, van Gerven, Kording, Richards, Kietzmann, Lindsay, & Kriegeskorte (Golan et al.)** commentaries. Many of the (over 20) authors have played leading roles in developing this new field comparing DNNs to humans, and in both commentaries, the authors are advancing research agendas going forward. However, the authors fail to address any of our concerns, and at the same time, mischaracterize some of our key positions.

R2. Do we recommend abandoning DNNs as models of human vision?

Many commentators claim that we are categorically rejecting DNNs as models of human vision (**Golan et al.**; **Hermann,**

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

Nayebi, van Steenkiste, & Jones [Hermann et al.]; **Love & Mok**; **Op de Beeck & Bracci**; **Summerfield & Thompson**; **Wichmann, Kornblith, & Geirhos [Wichmann et al.]**; **Yovel & Abudarham**), with quotes like:

> In the target article, Bowers et al. propose that psychologists should abandon DNNs as models of human vision, because they do not produce some of the perceptual effects that are found in humans (**Yovel & Abudarham**)
>
> Unlike Bowers et al. we do not see any evidence that future, novel DNN architectures, training data and regimes may not be able to overcome at least some of the limitations mentioned in the target article – and Bowers et al. certainly do not provide any convincing evidence why solving such tasks is beyond DNNs in principle, that is, forever (**Wichmann et al.**)
>
> Nevertheless, the target article advocates for jettisoning deep-learning models with some competency in object recognition for toy models evaluated against a checklist of laboratory findings (**Love & Mok**)
>
> …Bowers et al. take failures of ImageNet-trained models to behave in human-like ways as support for abandoning DNN architectures (**Hermann et al.**)

However, this is not our position. Indeed, in section 6.1 in the target article, we clearly lay out four different approaches to modelling that should be pursued going forward, the first of which is to continue to work with standard DNNs that perform well in identifying naturalistic images of objects but modify their architectures, optimization rules, and training environments to better account for key experimental results in psychology. This is exactly the view that so many commentators are endorsing. Nowhere in the target article do we advocate for "jettisoning" DNNs, and it is hard to understand why so many researchers claim that we have.

R3. Is image computability an entry requirement for developing models of human vision?

While we explicitly endorse a research programme that, amongst other things, compares image-computable DNNs to human vision (if severely tested), most of the commentators are less ecumenical and reject alternative modelling approaches in psychology and neuroscience that already account for some key aspects of human vision and the brain more generally. The main reason for this selective interest in DNNs is that only DNNs can recognize photographic images of objects at human or superhuman levels (under some conditions), that is, only DNNs are "image computable." This is considered an essential starting point for developing models of human vision (**Anderson, Storrs, & Fleming [Anderson et al.]**; **DiCarlo et al.**; **Golan et al.**; **Love & Mok**; **Op de Beeck & Bracci**; **Spratling**; **Summerfield & Thompson**; **Wichmann et al.**; **Yovel & Abudarham**). As Spratling puts it "… the ability to process images would seem to me to be a minimum requirement for a model of vision, and models that cannot be scaled to deal with images are not worth evaluating." Similarly, Summerfield & Thompson describe working with nonimage-computable models as "regressive." Not to be outdone, Love & Mok write:

> The authors invite us to return to the halcyon days before deep learning to a time of box-and-arrow models in cognitive psychology and "blocks world" models of language (Winograd, 1971), when modelers could narrowly apply toy models to toy problems safe in the knowledge that they would not be called upon to generalize beyond their confines nor pave the way for future progress.

This emphasis on image computability betrays a fundamental misunderstanding of what models are and what they are for. The goal of a scientific theory/model in the cognitive sciences is to account for capacities, predict data, and explain key phenomena, not to superficially resemble that which it purports to explain. When developing DNNs of human vision, image computability makes a system *look like* a visual system, but it does not make that system a good *model* of the human visual system. The ability to identify photorealistic images is a perk, not a barrier to entry. The barrier to entry is explanatory power and accounting for key empirical results. Rather than dismiss alternative approaches to modelling because they are not image computable, the relevant questions are "What have we learned from the multitude of modelling approaches available to vision scientists?" and "What are the most promising approaches going forward?"

To answer these questions, we need to consider the different modelling approaches of the past and the different approaches currently on offer. First, there is a long history in neuroscience and psychology of developing conceptual and mathematical theories of human vision that have provided insights into key empirical phenomena, from wiring diagrams designed to explain single-cell responses of simple and complex cells in V1 (Hubel & Wiesel, 1962), to dual-stream theories of vision designed to explain neuropsychological disorders of vision (Goodale & Milner, 1992), to theories of object recognition in normal vision (e.g., Biederman, 1987; Marr, 1982). These approaches to modelling are still active and providing valuable insights (Baker, Garrigan, & Kellman, 2021; Goodale & Milner, 2023; Vannuscorps, Galaburda, & Caramazza, 2021).

Second, there is a long history of building neural networks that process simple visual inputs to gain insights into the psychological and neural processes involved in object recognition, such as the neocognitron model (Fukushima, 1980) that implemented and extended the theory of Hubel and Wiesel, and the JIM model that implemented and extended the theory of Biederman (Hummel & Biederman, 1992). This latter model, JIM, and its successors (Hummel, 2001; Hummel & Stankiewicz, 1996) recognize simple line drawings of objects and are premised on the assumption that the goal of the ventral visual stream is to build a representation of the distal stimulus (the world and the objects in it) that can be used to understand the visual world. On this view, object classification is merely a consequence, not the be-all and end-all, of the ventral visual stream. Unlike current DNNs, JIM, and its successors account for many key psychological findings in human object recognition – such as the sensitivity of humans to part–whole relations – without being able to process naturalistic photographic images.

In a similar way, Grossberg et al. developed adaptive resonance theory (ART) models that quickly learn to classify simple visual patterns without forgetting past learning, that is, networks that solve the stability–plasticity dilemma (e.g., Carpenter & Grossberg, 1987; Grossberg, 1980). ART models not only account for a range of empirical findings reported in psychology and neuroscience (Grossberg, 2021), but they have also been used to solve engineering challenges (Da Silva, Elnabarawy, & Wunsch, 2019). Grossberg has also developed detailed models of low-level vision that take in simple visual inputs to capture a wide range of perceptual illusions (Grossberg, 2014). Expanding on the work of Grossberg, Francis, Manassi, and Herzog (2017) implemented networks that process simple visual inputs to explain a range of crowding phenomena that current DNNs cannot explain. In related work, George et al. (2017, 2020) developed recursive cortical networks that support the recognition of "captchas" and can account for several phenomena core to human vision, including some Gestalt phenomena (Lavin, Guntupalli, Lázaro-Gredilla,

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

Lehrach, & George, 2018). These models rely on segmentation and occlusion-reasoning in a unified framework to support object recognition, but only work with simple visual stimuli. These modelling efforts (and many others) largely fall into the second research programme we endorse in section 6.1 in the target article, namely, building networks that focus on explaining key psychological phenomena rather than image computability.

Third, there are active research programmes following the third approach we endorse in section 6.1 in the target article, namely, building models that support various human capacities that current DNNs struggle with (without focusing on the details of psychological or neuroscience research). But again, these models cannot process the photographic images that DNNs recognize. For example, Hinton, a coauthor of AlexNet, rejects current image-computable DNNs as models of human vision and is instead developing Capsule and GLOM models (Hinton, 2022; Sabour, Frosst, & Hinton, 2017). Hinton (2022) writes:

> There is strong psychological evidence that people parse visual scenes into part–whole hierarchies and model the viewpoint–invariant spatial relationship between a part and a whole as the coordinate transformation between intrinsic coordinate frames that they assign to the part and the whole [Hinton, 1979]. If we want to make neural networks that understand images in the same way as people do, we need to figure out how neural networks can represent part–whole hierarchies.

Indeed, current DNNs fail to represent objects in terms of their parts and relations even when explicitly trained to do so (Malhotra, Dujmović, Hummel, & Bowers, in press).

Similarly, generative models, such as variational autoencoders, are being developed that learn disentangled representations of visual elements of a scene (single hidden units that encode shape, colour, position, etc.; e.g., Higgins et al., 2016; Montero, Bowers, Ponte Costa, Ludwig, & Malhotra, 2022; Zhang et al., 2022) and object-centric learning models are being built to perform perceptual grouping (e.g., Anciukevicius, Fox-Roberts, Rosten, & Henderson, 2022; Locatello et al., 2020). To understand these principles, these models are frequently trained and tested on datasets of artificially created simple visual stimuli. **German & Jacobs** explicitly argue that variational autoencoders provide a more promising framework for understanding how human vision encodes objects in terms of their parts and relations between parts. But at present, exploring this requires working with simple rather than the photorealistic images.

The important point to emphasize here is that all these models would (and some actually do) receive low Brain-Scores (some cannot even be tested) because they cannot process the photorealistic inputs in ImageNet. Yet these models explore important phenomena in constrained settings. Are we supposed to discard these models because they cannot process and recognize photographs of objects? We think not. In our view, the diversity of modelling approaches in psychology (and the cognitive sciences more generally) fits well with the diversity of productive questions that can be asked about cognitive systems (cf., van Rooij, 2022). This is important to counteract the assumption that all worthwhile models of vision can recognize naturalistic photographs of objects or are on a trajectory towards becoming image computable.

R4. Are image-computable models the "current best" models of human vision

Still, it might be argued that image-computable DNNs that perform well on prediction-based experiments are the current best models of human vision because they provide more insights into human vision. However, we are struggling to see what the new insights are (although see our responses to **Anderson et al.** and **Op de Beeck & Bracci** below). Current DNNs account for few findings from psychology, and only do well on brain prediction-based studies when there is no attempt to rule out confounds as the basis of their successes. At the same time, DNNs that vary in terms of their architectures (CNNs vs. transformers), and objective functions (classification vs. image reconstruction) support similar levels of predictions on behavioural and brain benchmarks (e.g., Storrs, Kietzmann, Walther, Mehrer, & Kriegeskorte, 2021), with **Hermann et al.** and **Linsley & Serre** noting a recent trend for better performing models of object recognition doing more poorly on Brain-Score (although **Wichmann et al.** note that a transformer model trained on 4 billion images does much better on behavioural benchmarks). And as noted by **Xu & Vaziri-Pashkam**, when RSA is assessed with higher quality brain data, the correspondence across levels of DNNs and visual cortex is lost for familiar objects, and the predictivity scores go down dramatically for unfamiliar objects. More problematically, Xu & Vaziri-Pashkam note that RSA scores are greatly reduced following theoretically motivated experimental manipulations of images. What conclusions or insights about human vision follow from these observations? At present, it seems that the main advantage of image-computable DNNs compared to alternative models is that they recognize things, with little evidence that they do this in the way that humans do.

In fact, many commentators readily concede that current DNNs are doing a poor job in accounting for the results of experimental studies of human vision, and multiple possible solutions have been proposed. DNNs need to be trained with a better diet of images that more closely resemble human experience (**Linsley & Serre**; **Op de Beeck & Bracci**; **Yovel & Abudarham**), more biological constraints need to be added to models, such as representing binocular input from two eyes (**Chandran, Paul, Paul, & Ghosh**), and new objective functions and tasks need to be explored, including building DNNs that support vision for action (**German & Jacobs**; **Hermann et al.**; **Li & Mur**; **Liu & Bartolomeo**; **Rothkopf, Bremmer, Fiehler, Dobs, & Triesch**; **Slagter**; **Summerfield & Thomson**), with many of these authors advocating for some combination of the above approaches. Again, we agree with these research agendas, and we are pursuing some of these ourselves, including adding biological constraints to networks (Evans, Malhotra, & Bowers, 2022; Tsvetkov, Malhotra, Evans, & Bowers, 2023) and modifying training environments (Biscione & Bowers, 2022), in an attempt to make DNNs encode information in a more human-like manner. At the same time, there are good a priori reasons to think major architectural innovations may be necessary, for example, to encode relations between parts (**Kellman, Baker, Garrigan, Phillips, & Lu**), with some authors more pessimistic regarding the promise of DNNs as models of brains, with quotes such as: "Deep neural networks (DNNs) are not just inadequate models of the visual system but are so different in their structure and functionality that they are not even on the same playing field" (**Gur**) and the claim that DNNs "are doomed to be largely useless models for psychological research on language" (**Bever, Chomsky, Fong, & Piattelli-Palmarini [Bever et al.]**).

Of course, the human visual system is an image-computable neural network (although a network that differs from current DNNs in many fundamental ways; Izhikevich, 2004). However, the claim that current image-computable DNNs are the most

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

promising models of human vision going forward, despite the limited insights gathered thus far, is nothing more than a faith-based prophecy that may or may not pan out. In our view, researchers should be pursuing multiple different modelling approaches to advance our understanding of human vision. It is the dismissal of alternative approaches that is regressive (cf., Rich, de Haan, Wareham, & van Rooij, 2021, for a computational account of why this is detrimental).

R5. The role of prediction and explanation in model building

In the target article, we distinguished between uncontrolled, prediction-based studies that often highlight DNN–human similarities and controlled experiments that often highlight dissimilarities. We argued that the former experiments are problematic given that predictions can be driven by confounds whereas the latter experiments can help rule out confounds and allow researchers to draw causal conclusions regarding similarities and differences between DNNs and humans. To our surprise, few commentators even comment on this issue. The only exceptions are **Srivastava, Sifar, & Srinivasan** who highlight that similar issues apply in other domains, **Golan et al.** who highlight the importance of all variety of designs, and **Veit & Browning** who point out that properties and abilities of biological systems can be multiply realized and that controlled experiments are needed to make causal conclusions regarding the similarity of DNNs and humans.

Despite the potential problem of confounds in prediction-based studies, several commentators emphasize the importance of model predictions (**Golan et al.**; **Lin**; **Moldoveanu**; **Op de Beeck & Bracci**; **Veit & Browning**; **Wichmann et al.**; **Yovel & Abudarham**). For example, Wichmann et al. write: "we believe that both prediction and explanation are required: An explanation without prediction cannot be trusted, and a prediction without explanation does not aid understanding," and Lin writes "developing models with predictive accuracy might be a complementary approach that could help to test the relevance of explanatory models that have been developed through controlled experimentation."

These comments seem to suggest that testing models on controlled experiments does not involve prediction. In fact, both prediction-based studies and controlled experiments test model-based predictions (**Golan et al.**). The important distinction is between predictions with and without explanation. In the case of testing DNNs on prediction-based studies, there is no manipulation of independent variables designed to test specific hypotheses regarding how the models made their predictions, and accordingly, no explanation for any good predictions. Indeed, receiving 100% predictivity does not help the scientist understand how a DNN is predicting (see Fig. 5 in the target article). By contrast, in the case of testing DNNs on controlled experiments, the models are assessed in how well they predict performance across conditions designed to test hypotheses, and accordingly, good predictions can contribute to an explanation.

Of course, some types of predictions provide a stronger test of a model than others (**Spratling**), and this applies to both prediction-based studies and controlled experiments. In the case of prediction-based studies, current DNNs only perform well in the easy cases, namely, when training and test images are from the same distribution (often described as independent and identically distributed data or i.i.d. data). When DNNs are assessed on their ability to make behavioural or brain predictions for test images from a different distribution (out-of-distribution data or o.o.d. data), performance plummets. For example, as noted above, **Xu & Vaziri-Pashkam** showed that brain predictivity with RSA was much weaker when they included novel stimuli in the test set, and DNN successes on same-different visual judgements are limited to cases in which training and test images are similar (Puebla & Bowers, 2022, 2023). In other words, not only do prediction-based studies provide little insight into how models predict, but also their successful predictions are highly circumscribed.

Similarly, in the case of DNNs that successfully account for the results of controlled psychological experiments, the models predict that the controlled experiments will replicate on another sample of participants, images, and so on taken from the same population (i.i.d. data). But DNNs rarely make counterintuitive predictions that are subsequently confirmed in controlled experiments (analogous to predictions of o.o.d. data). It is worth noting that models tested on controlled experiments are generally described as *accounting for* (rather than *predicting*) results when successful, and this terminology might be more appropriate for prediction-based studies tested on i.i.d. data. Whatever the terminology, prediction-based studies and controlled experiments both assess how well DNNs predict (account) for data, but only the latter method tests hypotheses to rule out confounds and to make causal claims regarding how DNNs and humans identify objects.

Arguments regarding the relative advantages of prediction versus explanation touch on a broader debate regarding the relative advantages of studying natural systems in artificial conditions that allow precise control of variables versus naturalistic conditions where control is more limited. For example, **Love & Mok** cite the classic paper by Newell (1973) "You can't play 20 questions with nature and win" as a fundamental problem with studying the brain with controlled experiments. According to Love & Mok, laboratory studies in psychology have only produced a collection of findings they characterize as "cognitive science trivia." **Summerfield & Thompson** are not so dismissive of these experimental results, but they are critical of models in psychology that narrowly focus on explaining a small set of laboratory findings. DNNs, by contrast, are thought to hold promise of "genuine predictive power in the natural world" when trained on tasks that humans face in everyday life.

It strikes us as peculiar to characterize the empirical findings from psychology as "trivia" rather than core constraints for theory building and odd to dismiss models of specific empirical findings if they help explain key aspects of vision. What other area of science does not break down complex phenomena into parts? When **Summerfield & Thompson** highlight the narrow scope of psychological models with the example "…a model that explains crowding typically does not explain filling in and vice versa," it is important to note that current DNNs account for neither result.

For the sake of argument, let us accept the claim that image-computable models provide the best way forward for addressing Newell's challenge. Nevertheless, it is still the case that only controlled experiments provide specific hypotheses about how to improve DNN–human correspondences. For example, controlled experiments highlighted specific limitations of current DNNs as models of human vision (e.g., relying too much on texture, etc.) leading to specific suggestions about how to address them (e.g., a generative rather than discriminative objective function may result in a model that encodes shape rather than texture; **German & Jacobs**). A research programme of training image-computable DNNs on naturalistic datasets without running

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

specific controlled experiments will simply lead to black-box models in which there is no understanding of how the model works, let alone whether the model learns similar representations to humans.

It is also important to recognize the challenges with working with naturalistic images even when relying on controlled studies. For example, Rust and Movshon (2005) argued for the importance of building theories of biological vision using artificial and simple stimuli. They pushed back on the view that the best way to understand vision was to probe the system with naturalistic images, writing:

> Implicit in this approach is the assumption that synthetic stimuli are in some way impoverished or "simplistic" and therefore somehow miss important features of visual response. The main – and in our view, crippling – challenge is that the statistics of natural images are complex and poorly understood. Without understanding the constituents of natural images, it is imprudent to use them to develop a well-controlled hypothesis-driven experiment.

Although these comments were made before the current interest in DNNs, it remains just as difficult to design well-controlled hypothesis-driven experiments using natural images now as it was then given the billions of features associated with images. As a result, DNNs trained on these images become liable to learning based on shortcuts (Geirhos et al., 2020) and confounds (Dujmović, Bowers, Adolfi, & Malhotra, 2023), making it difficult to interpret their mechanisms and internal representations.

Finally, it is important to emphasize that model predictions are not the only way to advance our understanding of natural systems. **Lin** gives the example of Darwinian evolution as a model that has explanatory power but limited predictive accuracy. We think the term theory rather than model is more appropriate here, but the critical point is that evolution explains existing data very well, and it would be silly to dismiss the theory because it does not make precise predictions going forward. This point generalizes to all areas of science, such that unimplemented theories of vision can provide important insights into human vision if they can provide an account of key existing findings. Indeed, simply running experiments that test hypotheses can be highly informative. Of course, formal modelling has an important role to play, but in all cases, the focus should be on explanation, not prediction.

R6. The marketing of DNNs as the current best models of human vision is impeding our progress in developing better models

When comparing DNNs to humans it is not enough to carry out controlled experiments, it is also important to emphasize both the similarity and differences. This involves not only correctly characterizing the results from both DNNs and humans, but also carrying out studies that attempt to falsify claims regarding DNN–human similarities. Indeed, 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 claim if and only if the claim is false in some relevant manner (for a detailed discussion of the importance of severe testing when comparing DNNs to humans, see Bowers et al., 2023).

However, this does not characterize standard practice in the field at present. Instead, there appears to be a bias towards highlighting similarities and downplaying differences. Indeed, **Tarr** notes that many of the strong claims regarding DNN–human similarities are best understood as marketing rather than serious scientific claims – and on his view, the problem rests with the *consumers* who take the hype (too) seriously. He writes a story of a fool buying a pig because he saw a brochure suggesting pigs could fly. It is an allegory – the person should not be so naïve to believe the marketing. Similarly, he cautions us to be smart consumers of science and not take strong claims regarding DNN–human similarity too seriously. He writes that DNNs are only "proxy models" of vision and writes: "I don't think there is much actual confusion that deep neural networks (DNNs) are 'models of the human visual system.'"

We imagine it would be hard for **DiCarlo et al.** and **Golan et al.** to agree with this conclusion given they both repeat the claim that DNNs are the best models of human vision. But more importantly, this marketing impacts the field in two general ways.

R6.1. Marketing and research practices

When looking for DNN–human similarities, there is little motivation to move away from prediction-based studies that can provide misleading estimates of similarities, little reason for researchers to carry out controlled studies that provide severe tests of these claims, and little interest from editors and reviewers in publishing studies that highlight DNN–human dissimilarities. Consistent with these claims, two commentators explicitly minimize the importance of falsification. **Tarr** writes: "…less handwringing about what current models can't do; instead, they should focus on what DNNs can do." Similarly, **Love & Mok** write: "…we do not share their enthusiasm for falsifying models that are a priori wrong and incomplete." Instead, Love & Mok advocate for a Bayesian approach to model evaluation, where the question is which model is most likely given the data. But model selection depends on which data are under consideration, and currently, too many fundamental psychological findings are ignored because DNNs do not capture them. If Bayesian methods were used to select models that account for psychological phenomena, then in many cases, nonimage-computable models would perform best.

Perhaps the above comments are anomalous, and **Golan et al.** are right to doubt a bias against falsification in the field. But in our experience, this attitude towards falsification is widespread. For example, see the following NeurIPS workshop talk by Bowers (2022) that provides multiple examples of reviewers and editors stating that falsification is not enough. Rather, it is necessary to find "solutions" to make DNNs more like humans to publish: https://slideslive.com/38996707/researchers-comparing-dnns-to-brains-need-to-adopt-standard-methods-of-science. Similar biases are well recognized in other fields. For example, it is analogous to a bias against publishing null results in psychology that is well understood to have led to many false conclusions (Simmons, Nelson, & Simonsohn, 2011).

R6.2. Marketing and (mis)characterizing research findings

There is another respect in which this marketing manifests itself, namely, weak or ambiguous findings are too often characterized as supporting strong conclusions. We gave multiple examples of this in the target article (e.g., Caucheteux, Gramfort, & King, 2022; Duan et al., 2020; Hermann, Chen, & Kornblith, 2020; Kim, Reif, Wattenberg, Bengio, & Mozer, 2021; Messina,

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

Amato, Carrara, Gennaro, & Falchi, 2021; Zhou & Firestone, 2019) and there are more examples from the current commentaries themselves. For instance, **de Vries, Flachot, Morimoto, & Gegenfurtner (de Vries et al.)** criticize us for claiming that colour and form are processed entirely separately in V1 and cite some studies of theirs that show that DNNs do a good job in capturing important features of human colour processing. We take the point that the strong claims by Livingstone and Hubel (1988) need to be qualified given subsequent work (e.g., Garg, Li, Rashid, & Callaway, 2019), but de Vries et al. mischaracterize their own findings. They claim that categorical perception of colour emerges as a function of training models to classify objects and note that this effect did not emerge in a DNN trained to distinguish artificial from human-made scenes (de Vries, Akbarinia, Flachot, & Gegenfurtner, 2022). However, as reported in Appendix 7 of de Vries et al. (2022), an untrained DNN also showed some degree of categorical perceptual effects as well. This latter finding substantially weakens the evidence for their claim that colour perception emerges as a consequence of learning to classify objects.

Similarly, **Love & Mok** criticize us for not "engaging with work that successfully addresses their criticisms," but the evidence they report do not support their conclusions. Love & Mok give two examples from their own lab. First, they describe the work of Sexton and Love (2022) who note that RSA and linear prediction methods of comparing DNNs to brains rely on correlations and write: "Just as correlation does [not] imply causation, correlation does not imply correspondence." We agree. The problem is in how they draw correspondence claims. The authors assess whether brain signals can causally drive object recognition in DNNs by substituting the response elicited in an internal layer of a DNN with (a linear transform of) the brain response elicited by the same visual stimulus. They find that the activities from brain regions do indeed drive DNN object recognition performance above chance levels and take this as evidence that the representations in DNNs and brain are similar.

However, there are both empirical and logical problems with their studies and the conclusions they draw. Empirically, as reported in the Supplemental materials (Fig. S10 and Table S3), when brain data are used to drive DNN object recognition, performance drops from ~80 to <10% in one experiment and from ~58 to <2% in the second experiment. This large drop in performance is problematic for their conclusion. More fundamentally, the observation that brain responses support (limited) object recognition in DNNs does not address the issue of confounds. Just as texture-like representations in DNNs might be used to predict shape representations in cortex (leading to good RSA or Brain-Scores in the absence of similar representations), it is possible that shape representations in cortex can be mapped to texture-like representations in DNNs to drive object recognition to a limited extent. That is, the (weak) causal link between brain activation and DNN object recognition does nothing to address our concern that good predictions do not imply similar representations. Just as correlations do not imply causation, causation does not imply correspondence.

Love & Mok also describe a study by Dagaev et al. (2023) that they claim addresses a problem identified by Malhotra, Evans, and Bowers (2020), namely, that DNNs are so susceptible to shortcut learning that they will classify the images from CIFAR10 based on a single-pixel confound. Their solution involved introducing a *too-good-to-be-true* prior during training – if an image could be classified successfully by a low-capacity network (which Dagaev et al. use as a shortcut detector), the image is down-weighted during training a full-capacity network. This way, the full-capacity network only learned on images that, Dagaev et al. claim, are less likely to contain shortcuts. While this method is certainly of interest for a machine-learning engineer, it is of limited relevance to a cognitive scientist and does not address the criticisms made by Malhotra et al. (2021). Firstly, if the shortcut is widely prevalent in the dataset – in Malhotra et al. a diagnostic pixel was present in 80–100% of images – this method would fail. Secondly, there is nothing to say that shortcuts picked up by DNNs are necessarily easier to pick up by a low-capacity network. There could be many complex shortcuts, involving a conjunction of features that will be ignored by humans and picked up by full-capacity DNNs, but not by low-capacity DNNs. The point that Dagaev et al. miss is that we do not want models to ignore simple diagnostic visual features (humans rely on heuristics across a wide range of domains) but that they should learn *the right kind of* features, that is, models should incorporate appropriate human inductive biases, not whatever the low-capacity DNN does not happen to find diagnostic.

Yovel & Abudarham describe how DNNs capture the face-inversion effect, writing: "Interestingly, a human-like face inversion effect that is larger than an object inversion effect is found in DNNs." In fact, as shown by Yovel, Grosbard, and Abudarham (2022) and others, DNNs show similar size-inversion effects for face and nonface stimuli when trained with an equal number of images per category (e.g., when trained to identify the same number of human faces and birds of the same species). That is, the models showed an expertise inversion effect, not a face-specific inversion effect. This contradicts the bulk of current empirical evidence showing that humans exhibit a greater inversion effect for faces compared to other categories even when they are expert at the other category. To reconcile these findings with the modelling work, Yovel et al. (2022) argue that bird watchers are more expert at human faces compared to birds, and this is why they show larger face inversion effects. Future work may well support this hypothesis, and if so, it would provide a good example of DNNs explaining important psychological data. However, as it stands, the DNN results are inconsistent with most psychological data.

This is not to say that there are no examples of DNNs doing a good job at accounting for the results from controlled experiments. For instance, **Anderson et al.** describe the results of Storrs et al. (2021) who identified conditions in which DNNs do and do not replicate illusions of gloss in humans. They found that unsupervised but not supervised learning produced human-like results and suggest unsupervised learning may play a similar role in humans. Similarly, **Op de Beeck & Bracci** describe the controlled studies by Kubilius, Bracci, and Op de Beeck (2016) showing that DNNs trained on ImageNet are sensitive to many of the nonaccidental features described by Biederman (1987), a finding we found surprising but subsequently replicated in unpublished work.

However, these successes are, in our view, the exception, not the rule. A combination of relying so heavily on uncontrolled prediction-based studies, a bias against falsification in controlled studies, and selectively characterizing results to emphasize DNN–human similarities is not the way forward to advancing our understanding of human vision.

The same issues apply when large language models are also frequently compared to human language. In the target article, we gave the example of Caucheteux et al. (2022) making strong conclusions about human language despite the fact that the

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

DNNs accounted for approximately 0.004 of the BOLD variance in response to spoken sentences. Similarly, Schrimpf et al. (2021) report that transformer models predict nearly 100% of explainable variance in neural responses to written sentences and suggest that "a computationally adequate model of language processing in the brain may be closer than previously thought." However, the strong claims from the article are undermined from data reported in the appendices. From Appendix S1 one finds out that the explainable variance is between 4 and 10% of the overall variance in three of the four datasets they analyse, and from the Appendix section "SI-1 – Language specificity," we find out that DNNs not only predict brain activation of language areas, but also nonlanguage areas, and in some analyses, the predictions are numerically larger for nonlanguage areas. Rather than providing evidence that these models process language like humans, the correlations may be more akin to the spurious correlation observed between mouse brain activations and cryptocurrency markets (Meijer, 2021).

Furthermore, as noted by **Houghton, Kazanina, & Sukumaran (Houghton et al.)**, when a child is learning to speak, it is unlikely that she is focusing on predicting the next word. Rather, it seems likely that she is trying to communicate thoughts and desires. That is, these models learn to produce well-formed syntactic sentences when trained on arguably the wrong objective function. Similarly, these DNNs do not appear to share human-like inductive biases in learning languages, what **Bever et al.** call a universal grammar. These innate properties of humans allow the child to learn languages with many orders of magnitude less training than DNNs (human learning must be compatible with the poverty of the stimulus constraint), and at the same time, limits the types of languages that the human language system acquires (unlike language learning in DNNs; Mitchell & Bowers, 2020). In our view, research with DNNs in the domain of language provides another example that good predictions in uncontrolled studies provide little evidence that DNNs rely on human-like representations, processes, or even objective functions.

We do agree with **Houghton et al.** that it can be useful to compare language in DNNs and humans to explore the capacities of DNNs that do not have any language-specific learning mechanism. But at present, not only do the learning objectives and learning constraints seem wildly different in the two systems, but also, the performance of fully trained models "sharply diverges" from humans in controlled experiments (Huang et al., 2023).

R7. The Brain-Score neuroconnectionists

Before concluding, we thought it would be worthwhile to focus on the commentaries by **DiCarlo et al.** and **Golan et al.**. Many of these authors have been amongst the most vocal in highlighting DNN–human similarities, and in both commentaries, they are describing agendas for how to push the field forward.

Perhaps most surprising for us, **DiCarlo et al.** do not even attempt to address the core problem with prediction-based studies used in Brain-Score, namely, predictions of observational datasets might be mediated by confounds. Instead, they mischaracterize our views regarding benchmarks, writing:

> Bowers et al. eschew community-transparent suites of benchmarks yet they imply an alternative notion of vision model evaluation, which is somehow not a suite of benchmarks… we see no alternative to support advances in models of vision other than an open, transparent, and community-driven way of model comparison.

Where **DiCarlo et al.** get the impression that we are opposed to "open, transparent, and community-driven way of model comparison" is beyond us. Rather, we caution against prediction-based studies and endorse controlled experiments to assess models, including image-computable DNNs. Indeed, we are building our own (open, transparent, and community-driven) evaluation suite, that we call *MindSet*, that will make it easy for researchers to assess image-computable DNNs against key findings in psychology (Biscione et al., 2023). MindSet facilitates the testing of DNNs across a series of controlled psychological experiments, each of which tests a specific hypothesis regarding how DNNs process and represent information.

The authors also report on an upcoming update on Brain-Score, with the inclusion of a controlled study by Baker and Elder (2022). They note that some DNN vision models tested on this dataset are within the noise ceiling of human data. It will be interesting to see these results given that Baker and Elder reported that VGG19, ResNet50, CorNET, and a visual transformer all failed to capture human results, writing:

> Our configural manipulation reveals an enormous difference in how humans and networks recognize the objects: while humans rely profoundly on configural cues, networks do not.

Regardless of how current DNNs perform on this specific dataset, we welcome the introduction of controlled studies to the Brain-Score benchmark. But if the authors of Brain-Score modify their benchmark to assess the results of controlled experiments, they will need to assess models in terms of how well they explain the impact of independent variables that test specific hypotheses rather than rank models by their overall prediction accuracy.

DiCarlo et al. also defend their claim that DNNs are the current leading models of human ventral visual processing and write: "Bowers et al. critique ANN models without offering a better alternative: They imply that better models exist or should exist, but do not elaborate on what those models are." They set the bar quite low for "best" given that current DNNs do extremely poorly in predicting the results of experiments that manipulate independent variables and provide little insight into how humans identify the objects included in current behavioural and brain benchmark studies. But in any case, we have detailed a long list of alternative models in section 6.1 in the target article in section R3 in our response. In our view, these nonimage-computable models have provided more insight into human vision thus far. Still, going forward, we do think it is important to try to build image-computable DNNs that do account for controlled studies, and in parallel, pursue alternative modelling approaches.

Golan et al. describe a progressive Lakatosian research programme they call "neuroconnectionism" (Doerig et al., 2023) that generates a rich variety of falsifiable hypotheses and advances through model comparison. They note that neuroconnectionism itself is best thought of as a computational language that cannot be falsified and that a failure of a specific DNN does not amount to a refutation of neural network models in general. The problem with this is that no one claims that a rejection of a specific model amounts to a falsification of DNNs in general, and no one rejects modelling as a core method for advancing science. They are mounting a defence against an imaginary critique (as do other

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

commentators, as noted in sect. R2). Our criticism with neuroconnectionism is that current claims regarding DNN–human similarity are grossly overstated because researchers rely too heavily on uncontrolled prediction-based studies and avoid severe testing of their hypotheses. When the right methods are employed – namely, controlled experiments as used in virtually all other areas of science – models account for few empirical findings of interest to vision researchers.

Unlike **DiCarlo et al.**, **Golan et al.** do note some of the advantages of controlled experiments and briefly touch on the limitations of uncontrolled prediction-based studies, writing:

> Controlled experiments pose specific questions. They promise to give us theoretically important bits of information but are biased by theoretical assumptions and risk missing the computational challenge of task performance under realistic conditions… Observational studies and experiments with large numbers of natural images pose more general questions. They promise evaluation of many models with comprehensive data under more naturalistic conditions, but risk inconclusive results because they are not designed to adjudicate among alternative computational mechanisms (Rust & Movshon, 2005). Between these extremes lies a rich space of neural and behavioral empirical tests for models of vision. The community should seek models that can account for data across this spectrum, not just one end of it.

But we do not find their arguments against controlled studies and in support of obsrevational studies persuasive. Yes, controlled studies are biased in the sense that they are driven by theoretical assumptions, but the unstated (and unknown) assumptions in uncontrolled studies do not avoid biased results. For example, the image datasets used in Brain-Score (see Fig. 2 in the target article) are not "neutral" and different results are obtained in other datasets (**Xu & Vaziri-Pashkam**). And what does it mean to claim that observational studies with naturalistic images promise to evaluate many models, and at the same time, note that this approach risks inconclusive results? Indeed, predictions made from naturalistic images taken from observational studies are, by their very nature, ambiguous as there are many potential confounds that can lead models to make predictions on the basis of shortcuts and confounds (Dujmović et al., 2023; Geirhos et al., 2020).

Furthermore, what does it mean to design tests that fall in-between observational and controlled studies? An experiment either does or does not manipulate independent variables designed to test hypotheses and rule out confounds. If the point is that it is important to work with image datasets that vary in their degree of complexity and naturalism, it remains the case that controlled experiments need to be run on all types of stimuli. Indeed, **Golan et al.** cite the discovery of texture bias and adversarial susceptibility as two examples of shortcomings of DNNs that have led to improvements. Putting aside the fact that current DNNs show almost none of the features of human shape processing and there are still no solutions to adversarial images, these limitations were both identified using controlled experiments that rely on complex but unnatural stimuli. Golan et al. do not identify any insights that have derived from uncontrolled studies.

Golan et al. also caricature psychology, writing: "Traditional psychological experiments are designed to test verbally defined theories." In fact, controlled experiments have been used to assess computational models in psychology long before the invention of AlexNet (e.g., Grossberg, 1967; Hummel & Biederman, 1992; Medin & Schaffer, 1978; Ratcliff & McKoon, 2008; Rescorla & Wagner, 1972; Shepard, 1987). This general lack of regard for formal models and results in psychology (not to mention the lack of regard for verbal theories) is impeding progress in characterizing DNN–human similarities and building better models of vision and the brain more generally. Indeed, this common and unwarranted attitude towards psychology partly motivated us to write the target article in the first place.

Golan et al. also defend the claim that DNNs are the "best models" of human vision, writing:

> The empirical reason why ANNs can be called the "current best" models of human vision is that they offer unprecedented mechanistic explanations of the human capacity to make sense of complex, naturalistic inputs.

Here perhaps we should take the advice of **Tarr** and appreciate this is more marketing than a scientific statement.

References

Anciukevicius, T., Fox-Roberts, P., Rosten, E., & Henderson, P. (2022). Unsupervised causal generative understanding of images. *Advances in Neural Information Processing Systems*, *35*, 37037–37054.

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

Baker, N., Garrigan, P., & Kellman, P. J. (2021). Constant curvature segments as building blocks of 2D shape representation. *Journal of Experimental Psychology: General*, *150* (8), 1556–1580.

Biederman, I. (1972). Perceiving real-world scenes. *Science (New York, N.Y.)*, *177*, 77–80.

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

Biscione, V., & Bowers, J. S. (2022). Learning online visual invariances for novel objects via supervised and self-supervised training. *Neural Networks*, *150*, 222–236.

Biscione, V., Yin, D., Malhotra, G., Dujmović, M., Montero, M., Puebla, G., … Bowers, J. S. (2023). Introducing the MindSet benchmark for comparing DNNs to human vision. *PsyArXiv*. https://doi.org/10.31234/osf.io/cneyp

Bowers, J. S. (2022). Researchers comparing DNNs to brains need to adopt standard Methods of Science. Invited workshop talk at *Neural Information Processing Systems*, New Orleans.

Bowers, J. S., Malhotra, G., Adolfi, F. G., Dujmović, M., Montero, M. L., Biscione, V., … Heaton, R. F. (2023). On the importance of severely testing deep learning models of cognition. *PsyArXiv*, 1–34. https://doi.org/10.31234/osf.io/wzns2

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54–115.

Caucheteux, C., Gramfort, A., & King, J. R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*, 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*, 242–247.

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

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., & Love, B. C. (2023). A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, *166*, 164–171.

Da Silva, L. E. B., Elnabarawy, I., & Wunsch, D. C. II. (2019). A survey of adaptive resonance theory neural network models for engineering applications. *Neural Networks*, *120*, 167–203.

de Vries, J. P., Akbarinia, A., Flachot, A., & Gegenfurtner, K. R. (2022). Emergent color categorization in a neural network trained for object recognition. *eLife*, *11*, e76472. https://doi.org/10.7554/eLife.76472

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., … Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*, 431–450. https://doi.org/10.1038/s41583-023- 00705-w

Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C. P., Lerchner, A., & Higgins, I. (2020). Unsupervised model selection for variational disentangled representation learning. In *Proceedings of the 8th international conference on learning representations*. https://openreview.net/forum? id=SyxL2TNtvr

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2023). Obstacles to inferring mechanistic similarity using Representational Similarity Analysis. *bioRxiv*. https://doi.org/10.1101/2022.04.05.487135

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/10.1017/S0140525X22002813 Published online by Cambridge University Press

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review*, *124*, 483–504.
Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.
Garg, A. K., Li, P., Rashid, M. S., & Callaway, E. M. (2019). Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science (New York, N.Y.)*, *364*(6447), 1275–1279. https://doi.org/10.1126/science.aaw5868
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*, 665–673.
George, D., Lazaro-Gredilla, M., Lehrach, W., Dedieu, A., & Zhou, G. (2020). A detailed mathematical theory of thalamic and cortical microcircuits based on inference in a generative vision model. bioRxiv, 2020-09.
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 (New York, N.Y.)*, *358*(6368), eaag2612.
Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*, 20–25.
Goodale, M. A., & Milner, A. D. (2023). Shape perception does not require dorsal stream processing. *Trends in Cognitive Sciences*, *27*, 333–334. https://doi.org/10.1016/j.tics.2022.12.007
Grossberg, S. (1967). Nonlinear difference-differential equations in prediction and learning theory. *Proceedings of the National Academy of Sciences of the United States of America*, *58*, 1329–1334.
Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1–51.
Grossberg, S. (2003). Filling-in the forms: Surface and boundary interactions in visual cortex. In L. Pessoa & P. de Weerd (Eds.), *Filling-in* (pp. 13–37). Oxford University Press.
Grossberg, S. (2014). How visual illusions illuminate complementary brain processes: Illusory depth from brightness and apparent motion of illusory contours. *Frontiers in Human Neuroscience*, *8*, 854. https://doi.org/10.3389/fnhum.2014.00854
Grossberg, S. (2021). *Conscious mind, resonant brain: How each brain makes a mind*. Oxford University Press.
Hermann, K. L., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, *33*, 19000–19015.
Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., … Lerchner, A. (2017). Beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International conference on learning representations, Toulon, France*.
Hinton, G. (1979). Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, *3*, 231–250.
Hinton, G. (2022). How to represent part-whole hierarchies in a neural network. *Neural Computation*, *35*, 413–452.
Huang, K., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2023). Surprisal does not explain syntactic disambiguation difficulty: Evidence from a large-scale benchmark. *PsyArXiv*, 1–79. https://doi.org/10.31234/osf.io/z38u6
Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106–152.
Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, *8*, 489–517.
Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517. https://doi.org/10.1037/0033-295X.99.3.480
Hummel, J. E., & Stankiewicz, B. J. (1996). An architecture for rapid, hierarchical structural description. In T. Inui & J. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 93–121). MIT Press.
Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology*, *23*, 420–456.
Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, *15*(5), 1063–1070. https://doi.org/10.1109/TNN.2004.832719
Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, *4*, 251–263.
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*, e1004896.
Lavin, A., Guntupalli, J. S., Lázaro-Gredilla, M., Lehrach, W., & George, D. (2018). Explaining visual cortex phenomena using recursive cortical network. *bioRxiv*, 380048. https://doi.org/10.1101/380048
Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science (New York, N.Y.)*, *240*, 740–749.
Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M. (2020, November). Weakly-supervised disentanglement without compromises. In *International conference on machine learning, Vienna, Austria* (pp. 6348–6359).
Malhotra, G., Dujmović, M., Hummel, J., & Bowers, J. S. (in press). Human shape representations are not an emergent property of learning to classify objects. *Journal of Experimental Psychology: General*.
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. https://doi.org/10.1016/j.visres.2020.04.013
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*, 899–917.
Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207.
Meijer, G. (2021). Neurons in the mouse brain correlate with cryptocurrency price: A cautionary tale. *Peer Community Journal*, *1*, e29.
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.
Mitchell, J., & Bowers, J. (2020, December). Priorless recurrent networks learn curiously. In *Proceedings of the 28th international conference on computational linguistics* (pp. 5147–5158). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.451
Montero, M., Bowers, J., Ponte Costa, R., Ludwig, C., & Malhotra, G. (2022). Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, *35*, 10136–10149.
Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science (New York, N.Y.)*, *257*, 1357–1363.
Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), Visual information processing: Proceedings of the eighth annual Carnegie symposium on cognition, held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972. Academic Press.
Puebla, G., & Bowers, J. S. (2022). Can deep convolutional neural networks support relational reasoning in the same-different task?. *Journal of Vision*, *22*, 11. https://doi.org/10.1167/jov.22.10.11
Puebla, G., & Bowers, J. S. (2023). The role of object-centric representations, guided attention, and external memory on generalizing visual relations. arXiv preprint *arXiv:2304.07091*.
Ramachandran, V. S., & Gregory, R. L. (1991). Perceptual filling in of artificially induced scotomas in human vision. *Nature*, *350*, 699–702.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.
Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*, 618–660.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64–69). Appleton-Century Crofts.
Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43, No. 43).
Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*, 1647–1650. https://doi.org/10.1038/nn1606
Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, *30*, 3856–3866.
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 of the United States of America*, *118*, e2105646118.
Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*, eabm2219.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science (New York, N.Y.)*, *237*, 1317–1323.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.
Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*, 261–267.
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*, 2044–2064.
Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.
Tsvetkov, C., Malhotra, G., Evans, B. D., & Bowers, J. S. (2023). The role of capacity constraints in convolutional neural networks for learning random versus natural data. *Neural Networks*, *161*, 515–524.
Vannuscorps, G., Galaburda, A., & Caramazza, A. (2021). The form of reference frames in vision: The case of intermediate shape-centered representations. *Neuropsychologia*, *162*, 108053.
van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, *1*, 127–128. https://doi.org/10.1038/s44159-022-00031-5

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

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*, 1172.

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. AITR-235. Retrieved from http://hdl.handle.net/1721.1/7095

Yovel, G., Grosbard, I., & Abudarham, N. (2022). Computational models of perceptual expertise reveal a domain-specific inversion effect for objects of expertise. *PsyXiv*, 1–25.

Zeki, S. (1991). Cerebral akinetopsia (visual motion blindness). A review. *Brain*, *114*, 811–824.

Zhang, H., Zhang, Y. F., Liu, W., Weller, A., Schölkopf, B., & Xing, E. P. (2022). Towards principled disentanglement for domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8024–8034).

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

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