ANNUAL REVIEWS

Annual Review of Vision Science

Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception?

Felix A. Wichmann[1] and Robert Geirhos[2]

[1]Neural Information Processing Group, University of Tübingen, Tübingen, Germany; email: felix.wichmann@uni-tuebingen.de

[2]Brain Team, Google Research, Toronto, Canada; email: geirhos@google.com

Annu. Rev. Vis. Sci. 2023. 9:501–24

First published as a Review in Advance on March 31, 2023

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-120522-031739

Keywords

visual psychophysics, object recognition, deep learning, neural networks, computational models, computer vision

Abstract

Deep neural networks (DNNs) are machine learning algorithms that have revolutionized computer vision due to their remarkable successes in tasks like object classification and segmentation. The success of DNNs as computer vision algorithms has led to the suggestion that DNNs may also be good models of human visual perception. In this article, we review evidence regarding current DNNs as adequate behavioral models of human core object recognition. To this end, we argue that it is important to distinguish between statistical tools and computational models and to understand model quality as a multidimensional concept in which clarity about modeling goals is key. Reviewing a large number of psychophysical and computational explorations of core object recognition performance in humans and DNNs, we argue that DNNs are highly valuable scientific tools but that, as of today, DNNs should only be regarded as promising—but not yet adequate—computational models of human core object recognition behavior. On the way, we dispel several myths surrounding DNNs in vision science.

1. INTRODUCTION

Computational modeling in vision science has a long and rich history dating back at least to Schade's photoelectric analog of the visual system and Reichardt's motion detector (see Schade 1956, Hassenstein & Reichardt 1956, Reichardt 1957). Computational models are useful because, first, a thorough quantitative understanding of an aspect of human visual perception implies that we should be able to build a computational model of it. Second, a computational model can serve as a concrete, testable hypothesis of a theory and deepen our understanding through an iterative process of experimentation, model assessment, and model improvement.

Models in vision science come in many flavors, some borrowing computational elements such as filters or gain control from engineering, some using information theory to motivate or derive model properties, and others using Bayesian statistics to optimally read out the activity within a model to derive the model decision. There is substantial diversity of models in vision science, and these models typically use whichever method or algorithm promises to be most helpful in a given context (Cichy & Kaiser 2019).

In computer vision, a new class of algorithms from machine learning (ML), so-called deep neural networks (DNNs), have revolutionized the field. DNNs were explicitly developed to solve complex, real-world pattern recognition problems and are now capable of identifying objects in typical real-world photographs. DNNs entered center stage in 2012, when a DNN named AlexNet, designed and trained by Krizhevsky et al. (2012), comprehensively won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Having been trained on 1.2 million images, it was able to classify a large set of hitherto unseen images into 1,000 classes with an error rate of 16.4%—down from 25.8% in the previous year. Ever since 2012, the ILSVRC has been won each year by a DNN, and DNNs are the de facto standard for pattern recognition in machine learning and computer vision.[1]

In essence, a DNN is a nonlinear function approximator implemented as a very large collection of simple units that are connected to other units by variable connection strengths, the so-called weights. The function of the DNN arises from the collective action of units; their connectivity, as specified by the DNN architecture; the weights connecting units; and the nonlinear activation function of the units, i.e., how inputs to a unit are transformed into its output passed to other units. DNNs are sometimes said to be based on neuroscience, with units (neurons) connected to each other via variable weights (axons and dendrites with synapses of varying strength). The initial inspiration for neural networks did indeed come from neuroscience, but DNNs only embody a highly simplified version of real neurons and their rich dynamics, intricate connectivity, and dendritic processing complexity. The term neural network in the name of DNNs should thus be taken with more than a grain of salt (see Douglas & Martin 1991).

In a somewhat simplified historical account, DNNs (with four or more layers) could be regarded as the children of (three-layer) connectionism popular from the mid-1980s to the late 1990s (Rumelhart et al. 1986, McClelland et al. 1986) and as the grandchildren of the (two-layer) perceptron popular from the late 1950s to the late 1960s (McCulloch & Pitts 1943, Rosenblatt 1958).[2] A comprehensive overview of the more complex and multistranded historical development

[1]The ILSVRC has not been run since 2017, given that DNNs have basically solved the challenge: State-of-the-art (SOTA) top-5 performance has less than 1.0% error (top-5 error rates indicate whether the ground truth label was contained in the top five model predictions). However, the ImageNet data set still serves as a benchmark and is widely used as such.

[2]As we state, this is an oversimplified DNN history; we are aware of some of the other DNN milestones like the Neocognitron (Fukushima 1980), LeNet (LeCun et al. 1989), or HMax (Riesenhuber & Poggio 1999), which cannot be placed conveniently in the family history presented above.

TOOLS VERSUS MODELS

- Tool: algorithm or statistical method that plays an important role in the scientific process but is not, in itself, of scientific interest; a means to an end
- Model: a concrete instance of a theory that is itself of scientific interest; different types of models exist, such as statistical or mechanistic models, as well as different modeling goals (see Section 3)

of deep learning is provided by Schmidhuber (2015). Accessible introductions to DNNs are provided, for example, by LeCun et al. (2015) and Kriegeskorte (2015). A comprehensive treatment of DNNs is provided by Goodfellow et al. (2016).

2. DEEP NEURAL NETWORKS AS TOOLS IN SCIENCE

DNNs are able to perform highly nonlinear mappings from potentially very high-dimensional inputs—such as images with tens of thousands of pixels—to potentially high-dimensional outputs—such as the 1,000 categories of ImageNet.[3] DNNs learn the nonlinear and high-dimensional input–output mapping from massive amounts of training data alone: They learn which features and dimensions of the input, and which of their nonlinear transformations and combinations, are relevant to solving a task.

This ability to find nonlinear, high-dimensional mappings makes DNNs a powerful tool in the sciences in general. In this review, we draw a distinction between tools and models—acknowledging that the distinction is somewhat blurry at its boundary (see the sidebar titled Tools versus Models). Tools are a means to an end and play an important role in the scientific process but are not, in themselves, of scientific interest. In general, a model, in contrast, is a concrete instance of a theory and is itself of scientific interest. Many statistical algorithms or methods and tests are tools. For example, we use linear regression or analysis of variance (ANOVA) for scientific inference but, typically, do not take them to be scientific models of the phenomenon under investigation.

As tools, DNNs excel in science. For example, DNNs have helped to speed up and improve single-molecule localization microscopy (Speiser et al. 2021), make fast and accurate 3D protein folding predictions (Senior et al. 2020), and substantially accelerate climate science simulations (Ramadhan et al. 2022). In applied mathematics, deep reinforcement learning was recently used to find faster matrix multiplication algorithms (Fawzi et al. 2022).

In the neurosciences and vision science, DNNs as tools have led to clear improvements of methods: DeepLabCut allows markerless pose estimation of single (Mathis et al. 2018) and multiple (Lauer et al. 2022) animals, helping in video-based observation and analysis of freely behaving animals. Simulation-based inference allows parameter inference in computational models that, without deep learning, would remain computationally intractable and has begun to be successfully applied to cognitive neuroscience (Gonçalves et al. 2020, Boelts et al. 2022). Goetschalckx et al. (2021) argue that so-called generative adversarial networks can be used to generate visual stimuli that are, on the one hand, complex and realistic but, on the other hand, offer much more control than stock images.

[3] In theory, shallow, three-layer networks are already universal function approximators (Hornik et al. 1989). In practice, however, given their finite number of hidden units and finite data sets and computing power, shallow networks did not succeed in solving many interesting, large-scale, or real-world problems.

However, the success of DNNs as tools goes beyond method improvements. An important additional role for DNNs is in exploration, as was clearly and convincingly argued for by Cichy & Kaiser (2019).[4] One aspect of exploration involves proof-of-concept or proof-of-principle demonstrations. Piloto et al. (2022), for example, showed that important aspects of intuitive physics can be acquired entirely through visual, bottom-up learning. In a similar vein in the domain of gloss perception, Storrs et al. (2021) used unsupervised DNNs to demonstrate how perceptual dimensions like gloss—only imperfectly corresponding to distal physical properties—can be learned from the entangled proximal stimulus without the need for (Bayesian) prior knowledge or generative models. DNNs can thus be used to explore how richly structured our visual environment is and how much about the (distal) 3D world could in principle be extracted from (proximal) 2D sensors in a purely discriminative fashion (see, e.g., Storrs & Fleming 2021).[5] Another exploratory aspect of DNNs is the generation of new hypotheses: In beautiful work, Rideaux et al. (2021) used a neural network to find a causal role for neurons in the macaque medial superior temporal area, whose tuning properties had hitherto been regarded as puzzling. Their neural network analysis suggests that these neurons may play a role in the decision of whether or not certain motion signals arise from the same source—and should thus be either combined or analyzed separately by downstream neurons, respectively. The usefulness of DNNs as tools in vision science is beyond doubt; their immense predictive power in high-dimensional input–output mappings is crucial for method development, for stimulus generation, and for exploration, as well as for proof-of-principle demonstrations and the generation and testing of new computational hypotheses.

It is sometimes tempting to turn successful statistical and computational tools into theories—the tools-to-theory heuristic (Gigerenzer 1991). Scientists often use the tools that they employ as metaphors for phenomena. An example mentioned by Gigerenzer is the idea of the mind as a statistician put forward by Brunswik after Pearson had developed inferential statistics. An example in vision science, where Bayesian statistics are often used as a tool to analyze experimental data, is the well-known Bayesian brain hypothesis, asserting that the visual system itself applies Bayesian probability calculus to sensory input (see Zednik & Jäkel 2016). Given the immense usefulness of DNNs as tools, and their success in object recognition in computer vision in particular, it is thus not surprising that they have also been proposed as computational models of human (core) object recognition in vision science (e.g., Yamins et al. 2014, Kriegeskorte 2015, Kubilius et al. 2019).[6] We evaluate DNNs as models of human core object recognition in Section 4, after discussing properties of good models.

3. WHAT MAKES A GOOD MODEL A GOOD MODEL?

As stated in Section 1, computational models are useful because, first, they force scientists to make assumptions explicit and specify dependencies within a model more precisely than is possible using

[4]We should note that we differ from Cichy & Kaiser (2019) in terminology but not in substance: We refer to DNNs for exploration as tools rather than models because they are typically used as a means to explore, similar to, e.g., a clustering algorithm or principal component analysis.

[5]One may speculate about how much Gibson would have appreciated DNNs as proof-of-principle tools, as he argued that the visual input alone—the optic array—is sufficiently rich to allow visual behavior (Gibson 1950), whereas many Bayesian or predictive coding accounts of vision stress the—allegedly—impoverished nature of the visual input, requiring prior knowledge and/or generative processes to accomplish visual perception. DNNs have clearly demonstrated that much more visual information can be obtained from images or videos alone than some vision scientists presume.

[6]In the case of DNNs, the tools-to-theories heuristic may even be a theory-to-tools-to-theory heuristic, as neural networks were initially inspired by the visual brain, and now DNNs are reimported into vision science (F. Jäkel, personal communication, Feb. 11, 2022).

language alone (see Brick et al. 2021). Second, based on prior knowledge and previous insights, they serve as concrete, experimentally testable hypotheses of theories and should explain data, processes, or how processes interact.

To model visual behavior, we traditionally employ mechanistic models, which capture human behavior in terms of inputs and outputs and whose computational ingredients are domain specific, i.e., informed by, or derived from, basic principles known and established in vision science. Typical mechanistic models of spatial vision, for example, employ spatial filters and divisive contrast gain control as some of their computational and causal building blocks (e.g., Goris et al. 2013, Schütt & Wichmann 2017). Mechanistic models in vision science can be classified into different types depending on how much they focus on behavior versus neurophysiological realism. Mechanistic models of psychophysics are abstracted away from the details of the biological implementation—without, hopefully, being outright neurally implausible. If neural plausibility or realism is one of the goals of modeling, then the models' ingredients or computational building blocks closely mimic the neuronal hardware and are linked to behavior by linking propositions (Teller 1984).

Statistical models, in contrast, are only concerned with fitting and predicting data, that is, with the correct mapping from inputs to outputs. Typically, they are generic and can be used in many different domains. DNNs used as tools, as described in Section 2, are a prime example of statistical models used as a means to an end only. However, some authors additionally claim DNNs to be (statistical) models of core object recognition (the ventral stream). Importantly, in the absence of any concrete computational domain knowledge, statistical models are often the only models that one could possibly use, and they thus often precede mechanistic models.

Different scientific goals lead to the use of different types of models, and there are also multiple modeling desiderata. For models of behavior, the following aspects are of particular importance: how well the model fits the data (predictivity) and how much the model aids understanding (explanation). In this section, we expand on these modeling desiderata and discuss how well, in general, DNNs as models fulfil them.

Mechanistic model: attempts to find a mechanistic, causal relationship between inputs and outputs that predicts the data; typically uses domain-specific components, processes, or transformations

Statistical model: probabilistic, general-purpose model concerned with fitting or predicting data and correct input–output mapping and describing the relationship between the independent (input) and dependent (output) variables

3.1. Predictivity

Traditionally, goodness of fit, i.e., how well the (already collected) data are described by the model, was assessed when modeling. Furthermore, if several models were compared, then model selection was applied to select the model offering the best trade-off between fitting the data and model complexity. With the advent of modern and highly adaptable ML classification algorithms, the focus shifted from explaining past data (goodness of fit) to a more stringent test, namely, predicting new data (often termed generalization in ML). DNNs are trained—their parameters are optimized—on training data, but their performance is assessed by how well they predict the previously unseen test data. DNN prediction performance is often spectacularly good, which is the reason why DNNs are currently so popular. Predictivity can be measured at different levels: at an aggregate level (e.g., whether models achieve human-level accuracy) but also at a more fine-grained level, for instance, asking whether models predict human response and error patterns at an individual stimulus or image level, which is a much stricter requirement (see Green 1964). In the case of classification data, predictivity at a fine-grained level can be measured by error consistency (Geirhos et al. 2020b), an image-level metric to assess the degree of similarity between, for instance, human and machine error patterns, i.e., whether humans and machines agree on which images are easy or difficult to classify (see also Rajalingham et al. 2018).

Good prediction performance, ideally at both the coarse- and the fine-grained level, is undoubtedly an important desideratum of a good model. However, we argue below that it is only one aspect of a good model. Using prediction performance on a specific task as a benchmark is exceedingly popular in computer vision and ML, and benchmarks are also gaining traction in

vision science. While benchmarks have fuelled a lot of progress, an overreliance on benchmarks to measure progress can be problematic if model quality is erroneously thought of as a single dimension along which models can be ranked, turning science into a spectator sport; as we argue in this review, model assessment is multidimensional, and a single number or rank does not do good models justice. Furthermore, benchmarks and rankings encourage small and short-term gains and discourage fundamental rethinking and the exploration of novel directions that, typically, are initially accompanied by worse performance on a benchmark.

3.2. Explanation

Beyond predictivity, an important desideratum of a good model is that it helps to explain the data and how a model's building blocks are causally linked to the observed behavior. Thus, a good model helps explain a scientific phenomenon, or at least aids in its interpretation. One way in which models can act as an explanation and aid in understanding is by being simple, i.e., having a limited number of parameters and containing building blocks or modules with identifiable subfunctions. Many of the mechanistic models in vision science are of this type; prime examples are the local motion-energy model (Adelson & Bergen 1985) and its extension to a global, medial-temporal, motion-processing model (Simoncelli & Heeger 1998). Such interpretable mechanistic models follow the adage of George Box:

> Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. (Box 1976, p. 792)

DNNs are made up of simple, well-understood computational units, but their decisions result from the interplay of hundreds of thousands of units and millions of connections. To date, DNN decisions remain largely opaque because the post hoc methods developed to understand DNNs—and thus provide an explanation for their behavior that is useful to a scientist—have not yet matured enough: Neither visualizing maximally activating features (Gale et al. 2020, Borowski et al. 2021, Zimmermann et al. 2021) nor heatmap or saliency methods (Kindermans et al. 2017, Montavon et al. 2017, Adebayo et al. 2018) sufficiently explain the functions learned by DNNs. While the methods may still mature, and there also exist promising novel approaches that may help in the future (e.g., Cohen et al. 2020, Chung & Abbott 2021), for now, DNNs do not provide as much of an explanation and do not aid in understanding as much as we would like them to—they should thus, at best, be regarded as statistical models in vision science.

It should not be left unmentioned, however, that it is not inconceivable that models of an organ as complex as the brain, and behaviors as complex as, for example, object recognition, cannot be modeled using a simple model—at least not with adequate prediction performance. Highly nonlinear systems may require highly nonlinear (statistical or mechanistic) models, and both types of models are notoriously difficult to understand and analyze. For successful models of visual perception, there may not be a dichotomy between understandable traditional (mechanistic) models on the one hand and impenetrable (statistical) DNNs on the other: Whether a model is understandable also depends on the complexity of the behavior that our model should predict. Thus, however much we like simplicity, it may well be that, if we want computational models of complex visual behavior, we have to come to terms with Santiago Ramón y Cajal's (1967, p. 240) insight: "Unfortunately, nature seems unaware of our intellectual need for convenience and unity, and very often takes delight in complication and diversity."

Ideally, models should be highly predictive, show human-like error patterns, and aid in understanding. In practice, all of these desiderata appear difficult to achieve when modeling complex

human visual behavior. Different models may be preferable in different situations depending on the modeling purpose: DNNs are preferable if a highly predictive model is desirable even if it does not provide much of an explanation. At other times, a simple, easy to understand mechanistic model that is not nearly as predictive may still be preferable. Thus, a model may be not only either good or bad, but also both: Depending on the purpose, models can simultaneously be great and poor or useful and futile. Model assessment is multidimensional.

4. ADEQUATELY MODELING VISUAL CORE OBJECT RECOGNITION

How adequate are current DNNs as models of human perception? We restrict ourselves to core object recognition (DiCarlo et al. 2012): our fast and effortless ability to classify objects in the real world or images (photographs) of objects as belonging to a particular class or category at the entry level—a cat, elephant, car, or house, etc. Clearly, human object recognition does more than classification (and thus more than core object recognition): We are also able to recognize individual exemplars, most obviously in face recognition (O'Toole & Castillo 2021) but also when recognizing our cat among a dozen other cats in the garden or our car in a parking lot (for a review, see, e.g., Logothetis & Sheinberg 1996). Obviously, there is also more to human vision than object recognition: We use vision to estimate properties of materials and scenes, assessing, for example, distances and surface angles, or to guide behavior and to navigate. However, notwithstanding how much else there is to vision, (core) object recognition is undoubtedly very important for perception:

> At a functional level, visual object recognition is at the center of understanding how we think about what we see. Object identification is a primary end state of visual processing and a critical precursor to interacting with and reasoning about the world. (Peissig & Tarr 2007, p. 76)

When assessing the adequacy of DNNs as models of human perception, it is fair and appropriate to use not only a task that is important for humans and computationally complex, but also one at which DNNs excel. In computer vision, object recognition is continuing to set the standards for DNN performance. The combination of these factors—the importance of object recognition for human perception, the central role of object recognition within computer vision, and the fact that DNNs trained on object recognition are being proposed as models for primate ventral stream core object recognition—renders visual core object recognition perhaps the best task for comparing human against machine behavior and thus assessing the adequacy of DNNs as models of a particularly important aspect of human visual perception.

What needs to be successfully modeled in core object recognition are the following central, often replicated, and most important findings (Biederman 1987): the (rapid) recognition of objects under changes in orientation and illumination, under partial occlusion, and if distorted by (moderate levels of) visual noise.

4.1. Robustness to 3D Viewpoints

Yamins et al. (2014) performed an impressive and influential set of experiments in which they optimized the architecture of their DNNs with respect to the performance of the DNN on an object recognition task. One of their main findings—not central to our discussion in this review—was a correlation between the performance of the DNNs on their object recognition tasks and the DNNs' ability to also predict neuronal firing patterns in the monkey inferior temporal cortex. What is central to the current discussion is that Yamins et al. varied the object views: from easy, canonical views in the low-variation condition to strong changes of the orientation of the objects in the high-variation condition. Human observers performed reasonably well across conditions

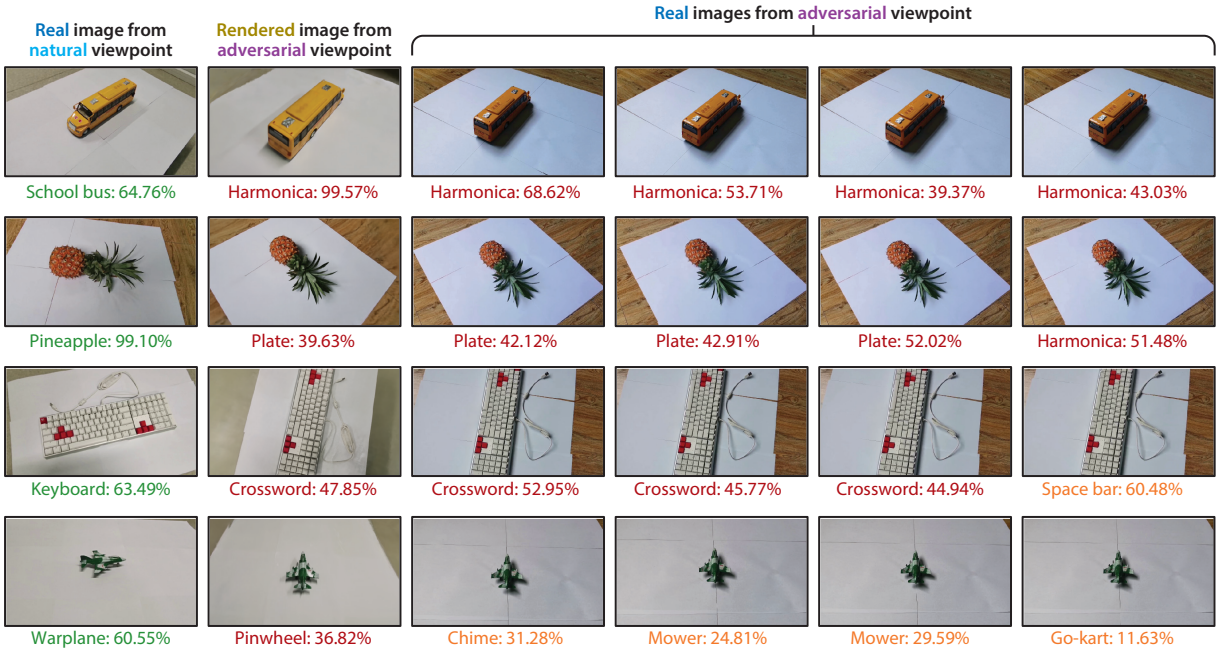

Figure 1

Visualization of 3D viewpoint dependence of deep neural network (DNN) object categorization. Column 1 shows the real images from natural viewpoints (correct classification; green label below images). Column 2 shows the rendered images from a viewpoint optimized to lead to wrong classification (adversarial viewpoints). Columns 3 to 6 show real images that approximate the adversarial viewpoints of column 2 (wrong classification; red or orange labels depending on DNN confidence in classification). Figure adapted with permission from Dong et al. (2023, figure 5).

(Yamins et al. 2014, figure 2*b*, p. 8621), since one of the strengths of human object recognition is its ability to cope with changes in orientation. Interestingly, their best DNN (resulting from hierarchical modular optimization) performed nearly as well, suggesting that DNNs may thus be on par with humans in their robustness to viewpoint or orientation changes.

However, the stimuli used by Yamins et al. (2014) were, first, comparatively few (eight exemplars from each of eight categories) and, second, not embedded in a natural background but instead superimposed. Dong et al. (2023) elegantly explored the viewpoint dependence of DNN object recognition systematically for several modern DNNs. For all of them, they found classification accuracy to be highly 3D view dependent: While classification accuracy was typically between 70% and 80% for standard ImageNet images, performance dropped dramatically with slightly unusual viewpoints to below 20% for most DNNs, with the best-performing DNN still below 50% accuracy.[7] Note that none of the viewpoints that are problematic for DNNs pose any difficulty for human observers: The images that are difficult for DNNs are quickly and effortlessly recognizable for us—at least in the examples shown in the paper (see **Figure 1**).

Similar results were obtained in related studies by Alcorn et al. (2019), Abbas & Deny (2022), and Ibrahim et al. (2022). Alcorn et al. concluded that their work "revealed how DNNs' understanding of objects like 'school bus' and 'fire truck' is quite naive—they can correctly label only a

[7] Dong et al. (2023) did not simply search for unusual viewpoints, but instead optimized them to find views that are difficult for DNNs, a procedure similar to how adversarial images are generated (see Section 4.4). For this reason, they call the viewpoints used adversarial viewpoints.

small subset of the entire pose space for 3D objects" (p. 4852). It is important to note that the images used by all of the studies cited above do not contain occlusions; the tested DNNs sometimes generalize poorly to novel 3D viewpoints despite seeing all of the relevant objects in full view. In this regard, a 3D viewpoint robustness gap between humans and machines remains.

4.2. Robustness to Image Distortions

Core object recognition performance should be not only (largely) viewpoint invariant, but also robust against other image distortions such as moderate levels of visual noise. Geirhos et al. (2018) systematically explored this issue for three DNNs (ResNet-152, VGG-19, and GoogLeNet) using 13 different image distortions or degradations. To several of the image distortions, DNNs were clearly much less robust than the psychophysically tested human observers; the discrepancy was particularly pronounced for uniform noise, low-pass and high-pass filtering, and the so-called Eidolon distortions (Koenderink et al. 2017) (see **Figure 2*a***). In the case of uniform noise, for example, human observers were still approximately 50–60% correct at a noise variance for which all tested DNNs were essentially at chance performance (6.25% in a 16-fold identification task). Including the image distortions in DNN training led to superhuman performance on the distortion included in the training, but there was little to no generalization to other distortions: Even training on undistorted images alongside seven different distortions did not help the DNNs to cope with previously unseen uniform noise. Thus, Geirhos et al. (2018) concluded that there is a large robustness gap between DNNs and human observers in their core object recognition ability in the face of image distortions, a conclusion that is corroborated by many other robustness studies (e.g., Berardino et al. 2017, Wichmann et al. 2017, Hendrycks & Dietterich 2019, Koh et al. 2021, Hendrycks et al. 2021a, Idrissi et al. 2022).

However, progress in deep learning is sometimes remarkably swift; thus, Geirhos et al. (2021) reassessed robustness to image distortions—termed out-of-distribution (OOD) robustness in ML—in 52 classic and SOTA DNNs in 2021 using a total of 17 OOD data sets and more than 85,000 psychophysical trials in the laboratory. Overall, different DNN architectures and training

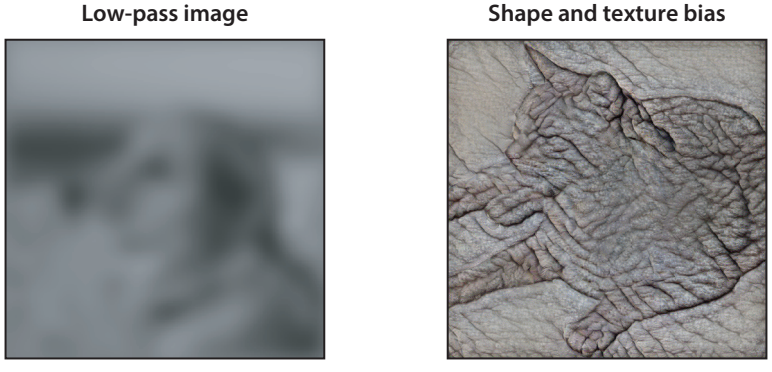

Figure 2

(*a*) Low-pass images are one of the remaining distortion types in which humans are currently still better than all 52 investigated diverse deep neural networks (DNNs) (Geirhos et al. 2021). Shown is a "dog" image at blur std = 7 pixels; human accuracy was 50–60%, and DNN accuracy was 10–40% averaged across many images at this blur level. Panel adapted with permission from Geirhos et al. (2018). (*b*) Shape bias is the tendency to classify images according to their global shape ("cat" in the example). Texture bias is the tendency to classify images according to their local texture-like characteristics ("elephant" in the example). Panel adapted with permission from Geirhos et al. (2019).

regimes appear to have little systematic influence on DNN robustness, but the sheer amount of training data did matter: DNNs trained on at least 14 million images showed near-human OOD robustness, with two models even surpassing humans in an aggregate measure of OOD robustness across the 17 data sets. However, none of the DNNs comes close to human robustness to low-pass filtering (see Geirhos et al. 2021, figure 2, p. 23891). We should also add that most of the DNN architectures popular in vision science—for example, AlexNet, VGG, ResNets, and Inception—showed poor and clearly subhuman OOD robustness if only trained on ImageNet. The lack of OOD robustness of these DNNs should be kept in mind if one compares one or some of them to human performance or uses them to derive neural similarity measures.

Furthermore, while increased OOD robustness through training on large-scale data sets is indeed an impressive achievement, we require more than just similar overall prediction performance if we want to assess DNNs as adequate models of human behavior: We also require error consistency even for only statistical models, as we argue in Section 3.1. With respect to error consistency, there remains a large gap between all of the 52 DNNs and human behavior: Whilst the 90 human observers showed large agreement in which images they felt were easy or difficult to recognize, human–machine error consistency is only at approximately half the value of human-to-human consistency even for the best DNNs (see Geirhos et al. 2021, figure 1*d*, p. 23890). Given the large remaining gap in error consistency between SOTA DNNs and human observers, it appears safe to conclude that even the highly OOD robust DNNs appear to process images differently from human observers.

4.3. Image Features Underlying Object Recognition

Initially, it was widely believed that DNNs recognize objects based on their shape—similar to how we believe humans recognize objects: "[T]he network acquires complex knowledge about the kinds of shapes associated with each category. . . . High-level units appear to learn representations of shapes occurring in natural images" (Kriegeskorte 2015, p. 429)—or that intermediate DNN layers recognize "parts of familiar objects, and subsequent layers. . .detect objects as combinations of these parts" (LeCun et al. 2015, p. 436).

However, systematic investigations now seriously challenge this view that DNNs, like humans, recognize objects via their shape. Baker et al. (2018) conducted a series of experiments in which DNNs (AlexNet and VGG-19) had to classify silhouettes of objects, silhouettes filled with the texture of another object, glass figurines, and outline figures. For all stimulus variations, DNNs performed poorly, suggesting that DNNs rely much more on surface characteristics like texture than do human observers and lack global shape sensitivity (see **Figure 2*b***). Similar results were obtained by Brendel & Bethge (2019), who successfully trained a ResNet variant they termed BagNet on ImageNet. BagNet exclusively relies on a bag of local features—with no shape encoding possible—but still performed similarly to several standard DNNs in terms of interactions between parts of the images, sensitivity to features, and errors.

While the two studies above assessed only DNN performance—either in response to changed stimuli or for a DNN architecture only capable of using local image patches for classification—Geirhos et al. (2019) compared human behavior directly with DNN classifications (DNN behavior) on cue-conflict stimuli created via style-transfer algorithms (Gatys et al. 2016), combining the shape of one object with the local surface characteristics (texture) of another, e.g., the shape of a cat with the texture (or skin) of an elephant. Human observers classified the cue-conflict stimuli almost exclusively according to their shape, whereas ImageNet-trained DNNs showed a strong texture bias, indicating that they do not classify objects according to their shape as humans do. Importantly, human–machine comparisons need to be careful and fair: To ensure a core

object recognition comparison, presentation time was limited to 200 ms (a single fixation), and all images were immediately followed by a high-contrast 1/f noise mask to minimize, as much as psychophysically possible, feedback influences on perception. Finally, observers had to respond quickly to ensure, again as much as possible, perceptual rather than cognitive responses, i.e., core object recognition.

The above result still held in 2021, when Geirhos et al. (2021) reassessed, among other performance measures, the shape and texture bias of 52 DNNs, including adversarial training, self-supervision, and image-text training [contrastive language-image pre-training (CLIP)]. Some modern DNNs—adversarially trained, CLIP—exhibited a stronger shape bias than previous DNNs, but even they remain substantially more texture biased than human observers (Geirhos et al. 2021, figure 3). A recent study by Malhotra et al. (2022) again confirms that humans, but not the DNNs investigated, have a shape bias even when learning novel stimuli for which texture is more predictive. Malhotra et al. argue that the human shape bias likely results from an inductive bias for shape (see Mitchell 1980, Zador 2019). Hermann et al. (2020) identified data augmentation as an important factor in increasing shape bias. Furthermore, Feather et al. (2019) generated synthetic DNN metamers: stimuli that are physically different but have indistinguishable activity at a certain layer of a DNN. While metamers for the early layers were recognizable by human observers, the activity of later-layer DNN metamers were not metameric for humans, indicating the existence of different representations of objects in DNNs and humans.

Given these results, it is safe to conclude that current DNNs typically do not use the same image features as humans do when recognizing objects. One particularly striking difference is that DNNs rely on local surface- or texture-like characteristics, whereas humans predominantly use shape.

4.4. Susceptibility to Adversarial Attacks

Szegedy et al. (2013) showed that, for standard DNNs, any image of category A can be made to be misclassified to belong to any arbitary category B through a tiny perturbation, often so small that it is invisible to the human eye.

4.4.1. Why are adversarial attacks problematic? Despite a decade of enormous research efforts to make DNNs robust against adversarial attacks, no principled defense mechanism has been found to date. The current best defense is brute-force adversarial training (Madry et al. 2017), which increases the perturbation needed to fool a DNN (an approach that may or may not achieve human-level robustness in the future). Adversarial examples are arguably among the most pressing open problems of deep learning research, and researchers have started to ask whether adversarial examples exist for human perception, too. Typical methods to find adversarials for DNNs cannot be applied to human perception, which is stochastic, is sequence dependent, and does not provide gradients. Despite these practical challenges, answering the question of whether there are adversarial examples for humans is highly relevant to the debate around DNNs as models of human visual perception (e.g., Dujmović et al. 2020), since their existence would indicate important similarities, whereas their absence would indicate important differences (see the sidebar titled Myth: Humans Suffer from Adversarial Vulnerability).

4.4.2. The crocodile conjecture: Humans don't suffer from adversarial examples. Do adversarial examples for humans exist? Strictly speaking, the answer is unknown, but we believe it to be exceedingly unlikely that proper adversarial examples exist for humans. Unfortunately, while adversarial examples do have a precise definition, other types of images have in recent years been

MYTH: HUMANS SUFFER FROM ADVERSARIAL VULNERABILITY

As explained in detail in Section 4.4, there are no known adversarial examples for the human visual system, at least not according to the proper definition used in ML. Of course, the visual system—like DNNs—is not error free: It can be shown to make wrong classifications if faced with hard images, ambiguous images, or visual illusions. However, such images are not adversarial images—they are hard, ambiguous, or visual-illusion images. All three kinds of images and their influence on human recognition are certainly worthy of study. However, progress in vision science is hindered by confusing such images with (proper) adversarial images and, even worse, claiming that there are deep similarities between DNNs and human vision because both allegedly suffer from adversarial vulnerability. The opposite is true: Precisely because only DNNs are known to be vulnerable to adversarial examples, there are substantial and important differences between DNNs and the human visual system in terms of how they process visual information.

described in some way or another as adversarial. In this section, we first precisely define what a convincing adversarial example for humans would be, applying the same definition that is used to define machine adversarials for a DNN f. Starting from an arbitrary image i with ground truth label l for which $f(i) = l$ holds (i.e., the original image is correctly classified by the model), an untargeted ϵ-adversarial image is an image $i + \delta$ such that $f(i + \delta) \neq l$ and $||\delta||_p \leq \epsilon$. This adversarial example is untargeted, since the perturbation δ (which is small according to some L_p norm, typically $p \in \{0, 1, 2, \infty\}$) just needs to fool the model into classifying $i + \delta$ as belonging to any class except the original one. In the case of a targeted adversarial, in contrast, a target class l' with $l' \neq l$ is chosen beforehand, and then any image $i + \delta$ such that $f(i + \delta) = l'$ is called a targeted adversarial example [subject to the same constraints as above, i.e., $f(i) = l$ and $||\delta||_p \leq \epsilon$].

Applying this to human visual perception would indicate that, if humans are indeed susceptible to ϵ-targeted adversarial examples, then for an arbitrary image i, such as the bananas in **Figure 3**, there would be a small perturbation δ ($||\delta||_p \leq \epsilon$) such that a human observer would classify $i + \delta$ as an arbitrary but prespecified other class, such as a crocodile. If ϵ is large enough, then this is trivially possible: We simply replace the banana image with a crocodile image. Thus, crucially, the perturbation bounded by ϵ needs to be small. A standard ResNet-50 model, for example, can be fooled into classifying the banana image into a baseball, power drill, or crocodile image with

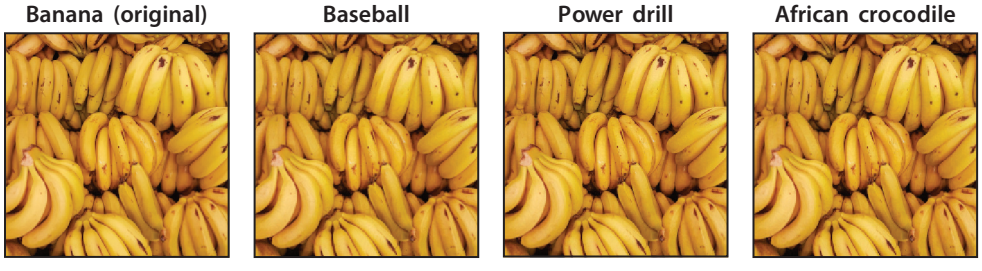

Figure 3

The crocodile conjecture: There are no adversarial examples for humans, i.e., it is impossible to make humans classify the bananas in the left image as a crocodile with a tiny perturbation. The original banana image (*left*) is adversarially modified (by adding tiny perturbations) such that a ResNet-50 then classifies the resulting perturbed adversarial images as a baseball, a power drill, or an African crocodile (depending on the perturbation). Results are based on the pretrained torchvision (Marcel & Rodriguez 2010) ResNet-50 implementation and the targeted L_∞ projected gradient descent attack with $\epsilon = 0.01$ implemented by Foolbox (Rauber et al. 2017). The original banana image taken from a photo by Rodrigo dos Reis on Unsplash, licensed under the Unsplash license (**https://unsplash.com/license**).

$\epsilon = 0.01$ (using the L_∞ norm and the [0, 1] image range), as shown in **Figure 3**. This means that, while every pixel may be changed, the maximum allowed perturbation is 0.01/1 or 1% for each pixel. For humans, we believe that it is impossible to find a perturbation δ subject to $||\delta||_\infty \leq 0.01$ that yields a crocodile classification for the banana image. We call this the crocodile conjecture (no adversarial examples for humans).[8] Given the many other types of images that have been termed adversarial, we believe that it is important to be precise when considering what a convincing adversarial example for human visual perception would constitute. In the following sections, we attempt to explain what adversarial examples are not: They are not hard images, nor ambiguous images, nor visual illusions.

4.4.3. What adversarial examples are not: hard images. To err is human, and making errors is a part of perception, whether biological or artificial. A popular image data set refers to natural, unmodified images that are particularly hard, i.e., images on which making errors is likely, as natural adversarial examples (Hendrycks et al. 2021b). While the data set is great, the name is not: These hard images have nothing to do with adversarial examples.

4.4.4. What adversarial examples are not: ambiguous images. Just as the existence of hard images is normal, so is the existence of ambiguous images. Examples include the famous rabbit-duck image (Jastrow 1899); hybrid images consisting of conflicting high and low spatial frequencies (Oliva et al. 2006); images that resemble both a cat and a dog, as in the work of Elsayed et al. (2018, figure 1 in versions 1 and 2, supplemental figure 1 in version 3); and multistable images (e.g., Safavi & Dayan 2022). Convincing adversarial examples are not tied to inherently ambiguous images—in fact, adversarial examples are problematic for DNNs precisely because they can be found for arbitrary images.

4.4.5. What adversarial examples are not: visual illusions. While the precise definition of visual illusions is subject to debate (Todorović 2020), it commonly involves a discrepancy between perception and reality. According to this definition, adversarial examples—if humans were susceptible to them—might be considered a new type of visual illusion. Importantly, however, the reverse is not true: None of the visual illusions that we know of are adversarial examples. Like ambiguous images, known visual illusions are nongeneral: Illusions are very particular stimuli or images, such as a particular configuration of shapes, as in the case of the Kanisza triangle; a particular configuration of luminances (reflectances), as in the case of lightness and brightness illusions; or a particular viewing angle, as in the case of the Necker cube. No (known) visual illusion matches the definition of an adversarial example.

4.4.6. Relevance of machine adversarials for humans. While no proper adversarials for humans are known, several studies investigated whether machine adversarials contain patterns that have relevance to humans, too. Zhou & Firestone (2019) showed that humans can sometimes anticipate the predictions of deep learning models in response to adversarial patterns—which is very different from being susceptible to adversarial patterns (for control experiments, see Dujmović et al. 2020). Furthermore, Elsayed et al. (2018) showed that human classification decisions at short presentation times (63–71 ms)[9] are influenced, to a certain degree, by

[8]We would love to be proven wrong; for anyone attempting to achieve this, a convincing demonstration would consist of a forced-choice paradigm where the choices are either banana or crocodile, with viewing times of at least 100–200 ms, foveal presentation, $\epsilon = 0.01$, and $p = \infty$.

[9]In addition, images were not shown in the central fovea (approximately 2 degrees of visual angle), but they were rather large, extending to 14.2 degrees of visual angle and thus into the periphery.

intermediate-size perturbations that are adversarial to deep learning models. Thus, intermediate-size (rather than small-size) adversarial perturbations are more than just arbitrary noise to humans. Unfortunately, this study is often erroneously interpreted as showing that there are adversarial examples for humans. The study's design choices purposefully target a regime where humans are highly prone to making errors (baseline classification performance for the choice between two classes is at approximately 75%, where 50% is chance level), and the study only used semantically related classes (dog/cat, cabbage/broccoli, spider/snake), not arbitrary ones. These are sensible choices for studying small effects under extreme conditions, but this is different from human core object recognition (DiCarlo et al. 2012). Elsayed et al. (2018) studied edge cases, using a paradigm that cannot test general perception for arbitrary classes (e.g., banana/crocodile) under conditions where the original image can be categorized with high accuracy.

Taken together, we believe that this evidence shows that humans do not have adversarials that can turn any arbitrary image (such as a banana) into an arbitrary other category (such as a crocodile) with only a tiny perturbation—i.e., the crocodile conjecture. Humans rely heavily on shape when recognizing objects, and object shape and boundaries cannot be significantly changed using tiny perturbations only. Adversarial examples are a profound problem for DNNs, but not for humans.

5. DISCUSSION AND OPEN ISSUES

5.1. Status Quo: Not Adequate, but Promising

As described in Section 4, in spite of their excellent prediction performance on standard image data sets like ImageNet, current DNNs see the world differently from human observers. Above, we review evidence that DNNs still lack robustness to changes in object pose (Section 4.1) and image distortions (Section 4.2); make nonhuman-like errors, as assessed by error consistency (Section 4.2); exhibit a lack of human-level shape bias (Section 4.3); and show a nonhuman susceptibility to adversarial images (Section 4.4). These behavioral differences can be exemplified in the following thought experiment: We could generate a data set containing only adversarial images and images with slightly different 3D viewpoints. While human classification performance would be largely unaffected, DNN performance, in contrast, could be driven to 0% correct. Conversely, given the low error consistency between DNNs and humans, we could select a subset of nonadversarially distorted images for which human performance is consistently poor but DNN performance is excellent. In effect, we could (almost) create a double dissociation between DNNs and human observers in terms of core object recognition performance—which would be impossible if both DNNs and humans recognized images similarly.

These profound behavioral differences indicate that current DNNs are not yet adequate behavioral models of human core object recognition. Nonetheless, we would like to stress that an assessment of the (in)adequacy of DNNs as models of human core object recognition behavior can only be a snapshot in time—it is true as of today. This does not mean that DNNs are forever, or for theoretical reasons, incapable of becoming adequate models of human core object recognition. In fact, for all of the current challenges, the tremendous progress in deep learning will very likely lead to improvements. In just over a decade, DNNs have come a long way from AlexNet, and they are likely to go much further still (see the sidebar titled Myth: Certain Tasks Cannot Ever Be Solved with Deep Neural Networks).

5.2. Excitement Versus Disappointment

Deep learning is often described as a revolution, and just as with any revolution, there is tremendous excitement and, simultaneously, profound disappointment. Vision science is no different: On the one hand, the ability of DNNs to fit neural data has led to the assessment that "deep

MYTH: CERTAIN TASKS CANNOT EVER BE SOLVED WITH DEEP NEURAL NETWORKS

Even though DNNs are getting better and better, there are still many tasks on which humans outperform them. It is therefore tempting to empirically investigate whether certain tasks cannot ever be solved with DNNs. We believe that such efforts are unlikely to succeed, for two reasons. First, DNNs are theoretically capable of solving (almost) any task, i.e., representing (almost) any input–output mapping, as they are universal approximators (Hornik et al. 1989). Second, training a specific network (or a handful of them) can never serve as a proof of nonexistence. This error is not infrequently witnessed when claims are made that a certain task cannot be solved by DNNs because a specific model cannot solve the task (given a certain data set, training regime, and objective function and a particular researcher's technical prowess). When investigating the question of whether a task cannot be solved by specific networks through a set of experiments, the results need to be contextualized, highlighting the necessarily restricted set of explorations and thus the necessarily restricted set of conclusions (context that often appears to be challenging to fit into a crisp paper title, abstract, or general summary). In contrast, theoretical proofs can sometimes allow very general statements to be made about a class of models. Thus, successfully training a specific DNN on some task can serve at least as a proof of concept. Failure to train a few specific DNNs on some task, in contrast, has limited implications.

hierarchical neural networks are beginning to transform neuroscientists' ability to produce quantitatively accurate computational models of the sensory systems" (Yamins & DiCarlo 2016, p. 364) or claims beyond core object recognition that "[d]eep neural networks provide the current best models of visual information processing in the primate brain" (Mehrer et al. 2021). On the other hand, behavioral shortcomings have also led to the assessment that there are "deep problems with neural network models of human vision" (Bowers et al. 2022). We believe that both of those perspectives are understandable, and that some—but not all—of these seemingly contradictory accounts can be reconciled through greater clarity about the goal of a particular model. Model quality is not a one-dimensional construct: Some models are good in some regards and poor in others. On the one hand, DNNs are the best predictive models in the history of core object recognition models—at least on standard computer vision benchmarks like ImageNet. On the other hand, currently, DNNs are also probably among the most inscrutable models, which constitutes a challenge to achieving the modeling goal of explanation.

5.3. Future Direction: Vision Science for Deep Learning

In Section 2, we discuss deep learning as a tool in science in general and vision science in particular. In Sections 4 and 5, we provide numerous examples of the value of vision science for deep learning: to understand how DNNs work, where they fail, and how they see the world. Psychophysical studies have revealed the current limitations of DNNs, which are of interest to those attempting to build not only better behavioral models, but also better deep learning models in general. Careful and fair comparisons (see Funke et al. 2021) are the hallmark of vision science, and the field, with its strong scientific foundation, has much to offer deep learning, which is still a predominantly engineering-driven area. In return, we may hope to develop better behavioral models of human visual perception, benefiting from rapid advances in deep learning and the field's engineering ingenuity (Ma & Peters 2020, Peters & Kriegeskorte 2021).

Interesting developments in the direction of better vision science–inspired methods for scrutinizing deep learning include generating controversial stimuli (Golan et al. 2020) to distinguish between candidate models (related to the idea of maximum differentiation competition; see Wang

MYTH: MORE (OF THE SAME) DATA IS ALL WE NEED

DNNs typically require in excess of one million images; one of the latest SWAG models by Singh et al. (2022) was trained on 3.6 billion images. While vision scientists recognize the crucial role of carefully controlled stimuli in experiments, large sets of images are impossible to curate on an image-by-image basis, and image data sets are harvested from the internet—with the hope that there is enough variation in scenes, objects, illuminations, poses, viewpoints, etc., such that there are, first, no systematic biases and, second, appropriately distributed (recognition) difficulties of the individual images.

Neither belief may be warranted, however: Meding et al. (2022) found image difficulty levels in ImageNet to be highly unbalanced, containing far too many trivial (and impossible) images. Furthermore, systematic biases in computer vision data sets typically do not disappear with data set scale, and ML models often exploit those biases. Examples include the spectral bias resulting from large-aperture portrait-style photos with shallow depth of field in rapid animal detection (Wichmann et al. 2010), as well as other cases of data set bias (Torralba & Efros 2011). Exploiting shortcuts in the data (Geirhos et al. 2020a) can lead to various generalization failures. One striking example is the 10–12% classification accuracy drop of DNNs when tested on a new ImageNet 2.0 data set created by Recht et al. (2019), who faithfully mimicked the original curation process: "This suggests that the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process. . . . It also shows that current classifiers still do not generalize reliably even in the benign environment of a carefully controlled reproducibility experiment" (Recht et al. 2019, p. 5389).

ML methods typically assume training and test sets to be independent and identically distributed (IID). For real-world natural images, what exactly constitute IID images is still an unresolved issue. Is a photograph of the same scene during a different season an independent image? Is one taken under different lighting, or one taken with the camera moved a little to one side or up or down? Are photos of a group of people taken from a different vantage point with different facial expressions independent images? Progress toward understanding these issues will likely help in making DNNs behave more like humans.

& Simoncelli 2008), using crowding measures to assess the importance of local versus global features (Doerig et al. 2020), and using a comparative biology approach when comparing human and machine visual perception (Lonnqvist et al. 2021), just to name a few. We are convinced that deep learning can benefit from vision science just as much as vision science can benefit from deep learning methods. This will be particularly true if we move beyond core object recognition to visual perception in general. Some of the challenges ahead are well articulated by Lake et al. (2017); a more sceptical position regarding DNNs as (only) models of perception and cognition is provided by Marcus (2018) [see the sidebar titled Myth: More (of the Same) Data Is All We Need].

5.4. Future Direction: Understanding Inductive Biases

In ML, the necessity of making assumptions to learn anything useful has been formalized by the no free lunch theorem (Wolpert & Macready 1997). However, understanding the inductive bias of a particular deep learning model—i.e., the set of assumptions that the model makes ahead of being exposed to data—is often incredibly challenging, and there are many complex interactions between, for instance, model architecture and data set. Nonetheless, we believe that it will be very important to improve our understanding of the assumptions that perceptual systems make. Humans are far from being a tabula rasa at birth (Zador 2019); we have a highly structured brain with appropriate inductive bias to help us learn rapidly and robustly from comparatively little data. In contrast, current DNNs still sometimes require more data than a human can possibly be exposed to during a lifetime (Huber et al. 2022).

There is much that remains to be understood. How do models and tasks interact with data sets? What are the inductive biases of biologically inspired filters (Dapello et al. 2020, Evans et al. 2022)? We observe that "recent work in AI. . .increasingly favoured computational architectures (for example, graph nets and tranformers) that implement an inductive bias towards relational, compositional processing" (Piloto et al. 2022, p. 1263), while standard convolutional networks are becoming less relevant. In a similar direction, Sabour et al. (2017) introduced capsule networks arguing that DNNs should have an inductive bias for explicit representation of geometric relationships of objects. As a final question, we should ask whether we require an explicit mid-level representation, the presemantic experience of the world (Nakayama et al. 1995, Anderson 2020). Currently, DNNs are trained end to end, from pixels to semantics in one model using one objective function. Should we explicitly train DNNs on psychologically inspired mid-level representations and then go from them to semantics? Building more adequate models of visual perception will likely require increasing attention to the various explicit and implicit assumptions, or inductive biases, made by different models (see the sidebar titled Myth: Recurrence Is Necessary to Solve Certain Tasks).

MYTH: RECURRENCE IS NECESSARY TO SOLVE CERTAIN TASKS

Given the important behavioral differences between DNNs and biological vision, one may wonder where these differences originate. One clear difference between brains and standard DNNs is that brains have recurrent connections, unlike standard feedforward DNNs. In line with this, several papers argue for the importance of recurrent processing for both biological and artificial systems (e.g., Serre 2019, Kietzmann et al. 2019, Kubilius et al. 2019, Kreiman & Serre 2020, van Bergen & Kriegeskorte 2020). Going beyond importance, however, the argument is sometimes made that recurrent DNNs are necessary to solve certain—challenging—tasks. In contrast, we do not think that recurrence is the key missing ingredient, since any algorithm that can be implemented by a recurrent DNN can also be implemented by a computationally equivalent feedforward DNN.

Since recurrent networks "once unfolded in time. . . , can be seen as very deep feedforward networks in which all the layers share the same weights" (LeCun et al. 2015, p. 442), there is no difference whatsoever between a finite time recurrent network and its unrolled feedforward counterpart in terms of what the model can compute (see also Liao & Poggio 2016). Any finite time recurrent network can be represented by a computationally equivalent finite depth feedforward network (e.g., via unrolling), and any infinite time recurrent network can be represented by a computationally equivalent infinitely deep feedforward network (keeping in mind that neither infinite time nor infinite depth would be particularly biologically plausible). Thus, there is no task that can only be solved with recurrence. For certain tasks, reusing weights may be useful—but this can be equivalently achieved by feedforward networks with weight sharing and by recurrent networks. In fact, in current DNN software libraries, most recurrent networks are trained as unfolded or unrolled feedforward networks with an algorithm called, tellingly, backpropagation through time—an implementational aspect well known by the researchers cited above and in the machine learning community in general. Any claim that recurrent networks have a computational advantage over feedforward networks—that is, that they can solve tasks that feedforward networks cannot solve—thus appears to be mistaken.

We believe that much of the debate around recurrence can be clarified when distinguishing between computation and implementation. As described above, any recurrent network can be unfolded into a (potentially very or even impractically deep) computationally equivalent feedforward network. At the implementational level, however, there are clear and important differences even between computationally equivalent recurrent and feedforward networks. For instance, artificial recurrent networks need less space to fit into memory, and biological ones need fewer neurons to fit into a brain and have obvious advantages in terms of energy efficiency and flexibility of processing. Whether one cares more about implementation or computation depends on the investigated question—but what is important is to distinguish between the two.

6. CONCLUSION

A decade ago, no researcher in vision science would have foreseen the phenomenal progress made by ML researchers in the field of neural networks. DNNs are spectacularly successful tools for (vision) science but also promising (statistical) models of core object recognition in terms of their prediction performance on standard computer vision image data sets. However, it is also fair to say that few researchers foresaw the substantial problems remaining for DNNs, despite their excellent prediction performance on standard computer vision image data sets: their lack of robustness to object pose and image distortions; their nonhuman-like errors, as assessed by error consistency; their still ill-understood dependence on the minutiae of the training images; their lack of human-like shape encoding; and their susceptibility to adversarial images. Last but not least, we are still lacking reliable tools to turn well-predicting but complex and nontransparent DNNs into human-understandable explanations—a desideratum of a scientific model. We argue, thus, that, as of today, DNNs should be regarded as promising—but not yet adequate—models of human core object recognition performance.

SUMMARY POINTS

1. Deep neural networks (DNNs) are powerful machine learning (ML) algorithms that have revolutionized computer vision and are increasingly important in vision research.
2. To assess their usefulness, it is important to be clear about goals and to distinguish between statistical tools and computational models. DNNs are great tools, but their usefulness as computational models in vision science is still subject to debate.
3. To become adequate computational models of human core object recognition, DNNs must use the same features as humans do and be able to robustly recognize objects despite variation in 3D viewpoints and image distortions, of which they are not yet capable.
4. Adversarial examples are images that have been carefully modified to fool a DNN into making a wrong prediction. No adversarial examples are known for humans; thus, the possibility of adversarial attacks on DNNs remains a major discrepancy between human and DNN visual perception.
5. It is unlikely that empirical investigations will be able to prove that certain tasks cannot ever be solved with DNNs because training a specific network can never serve as a proof of nonexistence.
6. There are no problems or tasks that can only be solved with recurrent DNNs, as any algorithm that can be implemented by a recurrent DNN can also be implemented by a computationally equivalent feedforward DNN.
7. Current DNNs should only be regarded as promising, but not yet adequate, computational models of human core object recognition behavior.

FUTURE ISSUES

1. The assumptions that DNNs make to learn, known as inductive biases, are often difficult to understand. It will be important to improve our understanding of these assumptions alongside the interactions among model architectures, tasks, and data sets to build more accurate models of human visual perception.

2. We are convinced that deep learning can benefit from vision science just as much as vision science can benefit from deep learning methods. Careful and fair comparisons are the hallmark of vision science, and the field, with its strong scientific foundation, has much to offer to deep learning, which is still a predominantly engineering-driven field.
3. Data sets—stimuli—matter enormously in vision science, and we have not yet sufficiently understood what makes natural images more or less similar to each other or more or less difficult to recognize, or what constitutes truly independent image sets required for training and testing in ML. Progress toward resolving these issues will likely lead to better DNN models of human visual perception, and data sets deserve at least as much attention as novel architectures.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank, in alphabetical order, Guillermo Aguilar, Bart Anderson, Blair Bilodeau, Wieland Brendel, Stéphane Deny, Roland Fleming, Tim Kietzmann, Been Kim, Thomas Klein, Pang Wei Koh, Simon Kornblith, David-Elias Künstle, Marianne Maertens, Sascha Meyen, Lisa Schut, Heiko Schütt, Kate Storrs, and Uli Wannek for helpful discussions and/or valuable feedback on aspects of the manuscript and Roland Zimmermann for foolbox advice. Furthermore, the authors are particularly indebted to Frank Jäkel and his numerous critical and insightful comments on previous versions of our manuscript. All opinions expressed in this article are our own and are not necessarily shared by any of the colleagues that we thank above. 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).

LITERATURE CITED

Abbas A, Deny S. 2022. Progress and limitations of deep networks to recognize objects in unusual poses. arXiv:2207.08034 [cs.CV]

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. 2018. Sanity checks for saliency maps. *Adv. Neural Inform. Proc. Syst.* 32:9505–15

Adelson EH, Bergen JR. 1985. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* 2(2):284–99

Alcorn MA, Li Q, Gong Z, Wang C, Mai L, et al. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4845–54. Piscataway, NJ: IEEE

Anderson BL. 2020. Mid-level vision. *Curr. Biol.* 30(3):R105–9

Baker N, Lu H, Erlikhman G, Kellman PJ. 2018. Deep convolutional networks do not classify based on global object shape. *PLOS Comput. Biol.* 14(12):e1006613

Berardino A, Laparra V, Ballé J, Simoncelli E. 2017. Eigen-distortions of hierarchical representations. *Adv. Neural Inform. Proc. Syst.* 30:3530–39

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94(2):115–47

Boelts J, Lueckmann JM, Gao R, Macke JH. 2022. Flexible and efficient simulation-based inference for models of decision-making. *eLife* 11:e77220

Borowski J, Zimmermann RS, Schepers J, Geirhos R, Wallis TSA, et al. 2021. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. In *Proceedings of the 2021 International Conference on Learning Representations (ICLR)*. N.p.: ICLR. **https://openreview.net/forum?id=QO9-y8also-**

Bowers JS, Malhotra G, Dujmović M, Montero ML, Tsvetkov C, et al. 2022. Deep problems with neural network models of human vision. *Behav. Brain Sci.* In press

Box GEP. 1976. Science and statistics. *J. Am. Stat. Assoc.* 71(356):791–99

Brendel W, Bethge M. 2019. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*. N.p.: ICLR. **https://openreview.net/forum?id=SkfMWhAqYQ**

Brick C, Hood B, Ekroll V, de-Wit L. 2021. Illusory essences: a bias holding back theorizing in psychological science. *Perspect. Psychol. Sci.* 17:491–506

Chung S, Abbott LF. 2021. Neural population geometry: an approach for understanding biological and artificial neural networks. *Curr. Opin. Neurobiol.* 70:137–44

Cichy RM, Kaiser D. 2019. Deep neural networks as scientific models. *Trends Cogn. Sci.* 23:305–17

Cohen U, Chung S, Lee DD, Sompolinsky H. 2020. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* 11:746

Dapello J, Marques T, Schrimpf M, Geiger F, Cox DD, DiCarlo JJ. 2020. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. bioRxiv 2020.06.16.154542. **https://doi.org/10.1101/2020.06.16.154542**

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 73(3):415–34

Doerig A, Schmittwilken L, Sayim B, Manassi M, Herzog MH. 2020. Capsule networks as recurrent models of grouping and segmentation. *PLOS Comput. Biol.* 16(7):e1008017

Dong Y, Ruan S, Su H, Kang C, Wei X, Zhu J. 2023. Viewfool: evaluating the robustness of visual recognition to adversarial viewpoints. *Adv. Neural Inform. Proc. Syst.* 37. In press

Douglas RJ, Martin KA. 1991. Opening the grey box. *Trends Neurosci.* 14(7):286–93

Dujmović M, Malhotra G, Bowers JS. 2020. What do adversarial images tell us about human vision? *eLife* 9:e55978

Elsayed G, Shankar S, Cheung B, Papernot N, Kurakin A, et al. 2018. Adversarial examples that fool both computer vision and time-limited humans. *Adv. Neural Inform. Proc. Syst.* 31:3910–20

Evans BD, Malhotra G, Bowers JS. 2022. Biological convolutions improve DNN robustness to noise and generalisation. *Neural Netw.* 148:96–110

Fawzi A, Balog M, Huang A, Hubert T, Romera-Paredes B, et al. 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* 610(7930):47–53

Feather J, Durango A, Gonzalez R, McDermott J. 2019. Metamers of neural networks reveal divergence from human perceptual systems. *Adv. Neural Inform. Proc. Syst.* 32:10078–89

Fukushima K. 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36(4):193–202

Funke CM, Borowski J, Stosio K, Brendel W, Wallis TSA, Bethge M. 2021. Five points to check when comparing visual perception in humans and machines. *J. Vis.* 21(3):16

Gale EM, Martin N, Blything R, Nguyen A, Bowers JS. 2020. Are there any "object detectors" in the hidden layers of CNNs trained to identify objects or scenes? *Vis. Res.* 176:60–71

Gatys LA, Ecker AS, Bethge M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–23. Piscataway, NJ: IEEE

Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, et al. 2020a. Shortcut learning in deep neural networks. *Nat. Mach. Intel.* 2:665–73

Geirhos R, Meding K, Wichmann FA. 2020b. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Adv. Neural Inform. Proc. Syst.* 33:13890–902

Geirhos R, Narayanappa K, Mitzkus B, Thieringer T, Bethge M, et al. 2021. Partial success in closing the gap between human and machine vision. *Adv. Neural Inform. Proc. Syst.* 34:23885–99

Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*. N.p.: ICLR. **https://openreview.net/forum?id=Bygh9j09KX**

Geirhos R, Temme CR, Rauber J, Schütt HH, Bethge M, Wichmann FA. 2018. Generalisation in humans and deep neural networks. *Adv. Neural Inform. Proc. Syst.* 31:7538–50

Gibson JJ. 1950. *The Perception of the Visual World*. Boston, MA: Houghton Mifflin

Gigerenzer G. 1991. From tools to theories: a heuristic of discovery in cognitive psychology. *Psychol. Rev.* 98(2):254–67

Goetschalckx L, Andonian A, Wagemans J. 2021. Generative adversarial networks unlock new methods for cognitive science. *Trends Cogn. Sci.* 25(9):788–801

Golan T, Raju PC, Kriegeskorte N. 2020. Controversial stimuli: pitting neural networks against each other as models of human cognition. *PNAS* 117(47):29330–37

Gonçalves PJ, Lueckmann JM, Deistler M, Nonnenmacher M, Öcal K, et al. 2020. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife* 9:e56261

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press

Goris RL, Putzeys T, Wagemans J, Wichmann FA. 2013. A neural population model for visual pattern detection. *Psychol. Rev.* 120(3):472–96

Green DM. 1964. Consistency of auditory detection judgments. *Psychol. Rev.* 71(5):392–407

Hassenstein B, Reichardt W. 1956. Systemtheoretische Analyse der Zeit, Reihenfolgen und Vorzeichenauswertung bei der Bewegungsrezeption des Rüsselkäfers Chlorophanus. *Z. Naturforsch. B* 11(9–10):513–24

Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, et al. 2021a. The many faces of robustness: a critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8340–49. Piscataway, NJ: IEEE

Hendrycks D, Dietterich T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*. N.p.: ICLR. **https://openreview.net/forum?id=HJz6tiCqYm**

Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15262–71. Piscataway, NJ: IEEE

Hermann K, Chen T, Kornblith S. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Adv. Neural Inform. Proc. Syst.* 33:19000–15

Hornik K, Stinchcombe M, White H. 1989. Multilayer feedforward networks are universal approximators. *Neural Netw.* 2(5):359–66

Huber LS, Geirhos R, Wichmann FA. 2022. The developmental trajectory of object recognition robustness: Children are like small adults but unlike big deep neural networks. arXiv:2205.10144 [cs.CV]

Ibrahim M, Garrido Q, Morcos A, Bouchacourt D. 2022. The robustness limits of SoTA vision models to natural variation. arXiv:2210.13604 [cs.CV]

Idrissi BY, Bouchacourt D, Balestriero R, Evtimov I, Hazirbas C, et al. 2022. ImageNet-X: understanding model mistakes with factor of variation annotations. arXiv:2211.01866 [cs.CV]

Jastrow J. 1899. The mind's eye. *Popul. Sci. Mon.* 54:299–312

Kietzmann TC, Spoerer CJ, Sörensen LK, Cichy RM, Hauk O, Kriegeskorte N. 2019. Recurrence is required to capture the representational dynamics of the human visual system. *PNAS* 116(43):21854–63

Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, et al. 2017. The (un)reliability of saliency methods. arXiv:1711.00867 [stat.ML]

Koenderink J, Valsecchi M, van Doorn A, Wagemans J, Gegenfurtner K. 2017. Eidolons: novel stimuli for vision research. *J. Vis.* 17(2):7

Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, et al. 2021. Wilds: a benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–64. N.p.: PMLR

Kreiman G, Serre T. 2020. Beyond the feedforward sweep: feedback computations in the visual cortex. *Ann. N. Y. Acad. Sci.* 1464(1):222–41

Kriegeskorte N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1:417–46

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Proc. Syst.* 25:1097–105

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, et al. 2019. Brain-like object recognition with high-performing shallow recurrent ANNs. *Adv. Neural Inform. Proc. Syst.* 32:12805–16

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017. Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253

Lauer J, Zhou M, Ye S, Menegas W, Schneider S, et al. 2022. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat. Methods* 19(4):496–504

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, et al. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1(4):541–51

Liao Q, Poggio T. 2016. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv:1604.03640 [cs.LG]

Logothetis NK, Sheinberg DL. 1996. Visual object recognition. *Annu. Rev. Neurosci.* 19:577–621

Lonnqvist B, Bornet A, Doerig A, Herzog MH. 2021. A comparative biology approach to DNN modeling of vision: a focus on differences, not similarities. *J. Vis.* 21(10):17

Ma WJ, Peters B. 2020. A neural network walks into a lab: towards using deep nets as models for human behavior. arXiv:2005.02181 [cs.AI]

Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. 2017. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 [stat.ML]

Malhotra G, Dujmović M, Bowers JS. 2022. Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Comput. Biol.* 18(5):e1009572

Marcel S, Rodriguez Y. 2010. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1485–88. New York: ACM

Marcus G. 2018. Deep learning: a critical appraisal. arXiv:1801.00631 [cs.AI]

Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, et al. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* 21(9):1281–89

McClelland JL, Rumelhart DE, Group PR, eds. 1986. *Parallel Distributed Processing*, Vol. II: *Explorations in the Microstructure of Cognition: Psychological and Biological Models*. Cambridge, MA: MIT Press

McCulloch WS, Pitts W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5(4):115–33

Meding K, Buschoff LMS, Geirhos R, Wichmann FA. 2022. Trivial or impossible—dichotomous data difficulty masks model differences (on ImageNet and beyond). In *Proceedings of the 2022 International Conference on Learning Representations (ICLR)*. N.p.: ICLR. **https://openreview.net/forum?id=C_vsGwEIjAr**

Mehrer J, Spoerer CJ, Jones EC, Kriegeskorte N, Kietzmann TC. 2021. An ecologically motivated image dataset for deep learning yields better models of human vision. *PNAS* 118(8):e2011417118

Mitchell TM. 1980. *The need for biases in learning generalizations*. Rutgers CS Tech. Rep. CBM-TR-117, Rutgers Univ., New Brunswick, NJ

Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* 65:211–22

Nakayama K, He ZJ, Shimojo S. 1995. Visual surface representation: a critical link between lower-level and higher-level vision. In *An Invitation to Cognitive Science*, ed. SM Kosslyn, DN Osherson, pp. 1–70. Cambridge, MA: MIT Press

Oliva A, Torralba A, Schyns PG. 2006. Hybrid images. *ACM Trans. Graph.* 25(3):527–32

O'Toole AJ, Castillo CD. 2021. Face recognition by humans and machines: three fundamental advances from deep learning. *Annu. Rev. Vis. Sci.* 7:543–70

Peissig JJ, Tarr MJ. 2007. Visual object recognition: Do we know more now than we did 20 years ago? *Annu. Rev. Psychol.* 58:75–96

Peters B, Kriegeskorte N. 2021. Capturing the objects of vision with neural networks. *Nat. Hum. Behav.* 5(9):1127–44

Piloto LS, Weinstein A, Battaglia P, Botvinick M. 2022. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nat. Hum. Behav.* 6(9):1257–67

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 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. *J. Neurosci.* 38(33):7255–69

Ramadhan A, Marshall JC, Souza AN, Lee XK, Piterbarg U, et al. 2022. Capturing missing physics in climate model parameterizations using neural differential equations. arXiv:2010.12559 [physics.ao-ph]

Ramón y Cajal S. 1967. The structure and connexions of neurons. In *Nobel Lectures, Physiology or Medicine 1901–1921*, ed. Nobel Found., pp. 220–53. Amsterdam: Elsevier

Rauber J, Brendel W, Bethge M. 2017. Foolbox: a python toolbox to benchmark the robustness of machine learning models. arXiv:1707.04131 [cs.LG]

Recht B, Roelofs R, Schmidt L, Shankar V. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–400. N.p.: PMLR

Reichardt W. 1957. Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems. *Z. Naturforsch. B* 12:447–57

Rideaux R, Storrs KR, Maiello G, Welchman AE. 2021. How multisensory neurons solve causal inference. *PNAS* 118(32):e2106235118

Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2(11):1019–25

Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65(6):386–408

Rumelhart DE, McClelland JL, PDP Res. Group, eds. 1986. *Parallel Distributed Processing*, Vol. 1: *Explorations in the Microstructure of Cognition: Foundations*. Cambridge, MA: MIT Press

Sabour S, Frosst N, Hinton GE. 2017. Dynamic routing between capsules. *Adv. Neural Inform. Proc. Syst.* 30:3856–66

Safavi S, Dayan P. 2022. Multistability, perceptual value, and internal foraging. *Neuron* 110(19):3076–90

Schade OH. 1956. Optical and photoelectric analogue of the eye. *J. Opt. Soc. Am.* 46:721–39

Schmidhuber J. 2015. Deep learning in neural networks: an overview. *Neural Netw.* 61:85–117

Schütt HH, Wichmann FA. 2017. An image-computable psychophysical spatial vision model. *J. Vis.* 17(12):12

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706–10

Serre T. 2019. Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* 5:399–426

Simoncelli EP, Heeger DJ. 1998. A model of neuronal responses in visual area MT. *Vis. Res.* 38(5):743–61

Singh M, Gustafson L, Adcock A, de Freitas Reis V, Gedik B, et al. 2022. Revisiting weakly supervised pretraining of visual perception models. arXiv:2201.08371 [cs.CV]

Speiser A, Müller LR, Hoess P, Matti U, Obara CJ, et al. 2021. Deep learning enables fast and dense single-molecule localization with high accuracy. *Nat. Methods* 18(9):1082–90

Storrs KR, Anderson BL, Fleming RW. 2021. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* 5:1402–17

Storrs KR, Fleming RW. 2021. Learning about the world by learning about images. *Curr. Direct. Psychol. Sci.* 30(2):120–28

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, et al. 2013. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]

Teller DY. 1984. Linking propositions. *Vis. Res.* 24(10):1233–46

Todorović D. 2020. What are visual illusions? *Perception* 49(11):1128–99

Torralba A, Efros AA. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–28. Piscataway, NJ: IEEE

van Bergen RS, Kriegeskorte N. 2020. Going in circles is the way forward: the role of recurrence in visual inference. *Curr. Opin. Neurobiol.* 65:176–93

Wang Z, Simoncelli EP. 2008. Maximum differentiation (MAD) competition: a methodology for comparing computational models of perceptual quantities. *J. Vis.* 8(12):8

Wichmann FA, Drewes J, Rosas P, Gegenfurtner KR. 2010. Animal detection in natural scenes: critical features revisited. *J. Vis.* 10(4):6

Wichmann FA, Janssen DH, Geirhos R, Aguilar G, Schütt HH, et al. 2017. Methods and measurements to compare men against machines. *Electron. Imaging Hum. Vis. Electron. Imaging* 2017(14):36–45

Wolpert DH, Macready WG. 1997. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1(1):67–82

Yamins DL, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19(3):356–65

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24

Zador AM. 2019. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* 10:3770

Zednik C, Jäkel F. 2016. Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese* 193(12):3951–85

Zhou Z, Firestone C. 2019. Humans can decipher adversarial images. *Nat. Commun.* 10:1334

Zimmermann RS, Borowski J, Geirhos R, Bethge M, Wallis TSA, Brendel W. 2021. How well do feature visualizations support causal understanding of CNN activations? *Adv. Neural Inform. Proc. Syst.* 34:11730–44