ANNUAL REVIEWS

Annual Review of Vision Science

The Quest for an Integrated Set of Neural Mechanisms Underlying Object Recognition in Primates

Kohitij Kar[1] and James J. DiCarlo[2]

[1]Department of Biology, Centre for Vision Research, York University, Toronto, Ontario, Canada; email: k0h1t1j@yorku.ca

[2]Department of Brain and Cognitive Sciences, MIT Quest for Intelligence, and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; email: dicarlo@mit.edu

Annu. Rev. Vis. Sci. 2024. 10:19.1–19.31

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

https://doi.org/10.1146/annurev-vision-112823-030616

Keywords

object recognition, ventral stream, inferior temporal cortex, artificial neural networks, neural mechanisms, visual intelligence

Abstract

Inferences made about objects via vision, such as rapid and accurate categorization, are core to primate cognition despite the algorithmic challenge posed by varying viewpoints and scenes. Until recently, the brain mechanisms that support these capabilities were deeply mysterious. However, over the past decade, this scientific mystery has been illuminated by the discovery and development of brain-inspired, image-computable, artificial neural network (ANN) systems that rival primates in these behavioral feats. Apart from fundamentally changing the landscape of artificial intelligence, modified versions of these ANN systems are the current leading scientific hypotheses of an integrated set of mechanisms in the primate ventral visual stream that support core object recognition. What separates brain-mapped versions of these systems from prior conceptual models is that they are sensory computable, mechanistic, anatomically referenced, and testable (SMART). In this article, we review and provide perspective on the brain mechanisms addressed by the current leading SMART models. We review their empirical brain and behavioral alignment successes and failures, discuss the next frontiers for an even more accurate mechanistic understanding, and outline the likely applications.

1. INTRODUCTION

Primates can rapidly infer and report multiple details about real-world objects in their field of view, despite the potentially infinite variation that an image of an object might present to the eyes (Rajalingham et al. 2015, 2018). How does this work?

A decade ago, experimental neuroscientists had already successfully probed the primate brain's visual processing pathways to identify a series of brain areas implicated in object identity and category inferences. These capabilities are commonly referred to as object recognition (for a review, see DiCarlo et al. 2012). In particular, prior work had demonstrated the central role of the ventral visual cortical stream for processing the visual input at the center of gaze to support object recognition behaviors (Ungerleider et al. 1982). In addition, neural recordings at the highest level of the primate ventral visual stream (Hung et al. 2005, Logothetis et al. 1995, Majaj et al. 2015) had demonstrated the neural population solution of primates' remarkable object recognition capabilities. Furthermore, anatomically constrained, specialized circuits had been discovered that exhibit selectively for specific visual objects and image statistics (Gross et al. 1972, Op de Beeck et al. 2001, Tanaka 1996). For instance, a population of neurons in the inferior temporal (IT) cortex that is more responsive to faces compared to other objects (Kanwisher et al. 1997, Tsao et al. 2006) was causally linked to face perception (Parvizi et al. 2012). Similarly, other studies have revealed different functional topographies in the ventral visual cortex (Lafer-Sousa & Conway 2013, Popivanov et al. 2014).

Despite decades of such experimental research, in 2013, our field did not know how object recognition worked. For example, we had not yet produced an end-to-end machine-computable model that could receive a two-dimensional pattern of photons striking the eye (i.e., an image) and transform this pixel-level information to accurately and rapidly solve visual object recognition. Similarly, we had not yet produced a machine-computable model that could accurately reproduce the ventral stream's intermediate series of steps in that solution (DiCarlo et al. 2012).

Working in parallel over decades, a small cadre of the computer vision community (LeCun & Bengio 1995, LeCun et al. 1989, Rumelhart et al. 1986) and the computational neuroscience community (Pinto et al. 2009, Riesenhuber & Poggio 1999) worked to stay close to the anatomy of the ventral stream. Beginning about a decade ago, machine vision system builders in this ventral stream–inspired lineage—fueled by more powerful computers and larger datasets (Russakovsky et al. 2015)—began to make remarkable strides in developing machine systems (He et al. 2016, Krizhevsky et al. 2012) that could solve the very hard problem of visual object recognition with near-human-level accuracy. At nearly the same time, visual neuroscientists began to show that these systems were by far the empirically leading scientific models of the primate brain mechanisms underlying object recognition (Cadieu et al. 2014; Khaligh-Razavi & Kriegeskorte 2014; Yamins et al. 2013, 2014).

This review is organized around reproducible models of the integrated set of neural mechanisms supporting object recognition; their evaluation, successes, and shortcomings; and how existing and future neural and behavioral data can help guide the development of the next generation of such models. While this review is focused on the nonhuman primate system, we think that this understanding will readily generalize to also explain the brain mechanisms of these same capabilities in humans (see the sidebar titled Rationale for the Nonhuman Primate Animal Model).

Before reviewing this progress, we define our premises: What do we mean by object recognition (Section 1.1)? What do we mean by an understanding of object recognition (Section 1.2)? More specifically, what is a mechanistic understanding of object recognition (Section 1.3)? Because object recognition is only a starting point, we synthesize our discussions around the goal of understanding the mechanisms of visual intelligence more generally.

RATIONALE FOR THE NONHUMAN PRIMATE ANIMAL MODEL

Why is this review focused on the rhesus macaque monkey's visual system? To successfully generalize the mechanisms of visual object recognition in humans, the animal model needs three things. First, it should have human-level perceptual capabilities (Rajalingham et al. 2015, 2018) (see **Figure 3**). Second, it should admit high-spatial- and -temporal-resolution neural measurements (Kar et al. 2019, Majaj et al. 2015) and targeted causal perturbations to interrogate brain circuits (Azadi et al. 2023, Kar & DiCarlo 2021, Rajalingham et al. 2021) that are not feasible in humans. Third, it should have evolutionary proximity (Perelman et al. 2011) and established brain-area homologies. The rhesus macaque meets all of these criteria [with some noted differences (see Rossion & Taubert 2019). Thus, an accurate model of the neural mechanisms of the monkey system will likely readily generalize to the homologous human brain system. In this review, when we use the word primate, we mean human and nonhuman primates. For more human-centric reviews, we refer the reader to Peters & Kriegeskorte (2021).

1.1. What Do We Mean by Object Recognition?

When a human observer encounters a visual scene, they quickly infer many things about the world content of that scene. To the extent that each report agrees with the underlying physical content of the world, we say that those inferences are accurate. The variables of the true physical content of the world, such as the number, shape, and category of objects and the position and pose of each object relative to the viewer, make up what is known as the latent content because the discrete or scalar values of these variables are not explicitly available to (i.e., they are hidden from) the perceptual system. For visual systems, such values must be inferred only from spatiotemporal patterns of photons striking the eyes. Remarkably, however, human reports of these values are highly accurate even with just single views of the scene—also known as single images. For instance, if you look briefly at the example image shown in **Figure 1**, you are likely able to answer many questions about the image: Did you see a bird or a cat? Did you see an owl or an osprey? Was the bird to the left or right of the fixation cross? Was the bird behind or in front of a branch? Was that a novel or a familiar bird? Is the bird pleasant or threatening? Typically, the study of object recognition within a scene is focused on primates' ability to determine the specific identity and category of a dominant foreground object (related to the first two questions) (**Figure 1*a***), but the broader domains of visual object perception and visual intelligence include all of these questions and many, many more.

To answer the question of whether we are making scientific progress on understanding the mechanisms (see Section 1.2) that underlie object recognition capabilities, it is essential to operationalize a starting set of tasks—both to assess and to characterize biological performance patterns and the performance patterns of computational models that aim to explain how that biology works. DiCarlo et al. (2012) proposed core object recognition as a starting point in that effort. By definition, core object recognition confines the visual intelligence challenge to the processing of images presented within the subject's central field of view (central 10° of visual angle) and for a limited time (<200 ms). This operational definition was chosen for three reasons: It is known that human shape discrimination abilities are best at the center of gaze, that the ventral visual stream processing is dominated by the central 10° (Op De Beeck & Vogels 2000, Ungerleider et al. 1982), that 200 ms corresponds to the duration of fixation during natural viewing behavior (DiCarlo & Maunsell 2000, Nuthmann 2017), and that object categorization performance at the center of gaze is remarkably accurate at this and even much shorter ones (Potter 1976, Thorpe et al. 1996).

Having rationally operationalized the sensory input domain above (10°, 200 ms), there are still many ways one might operationally assay the perceptual contents of human or animal minds

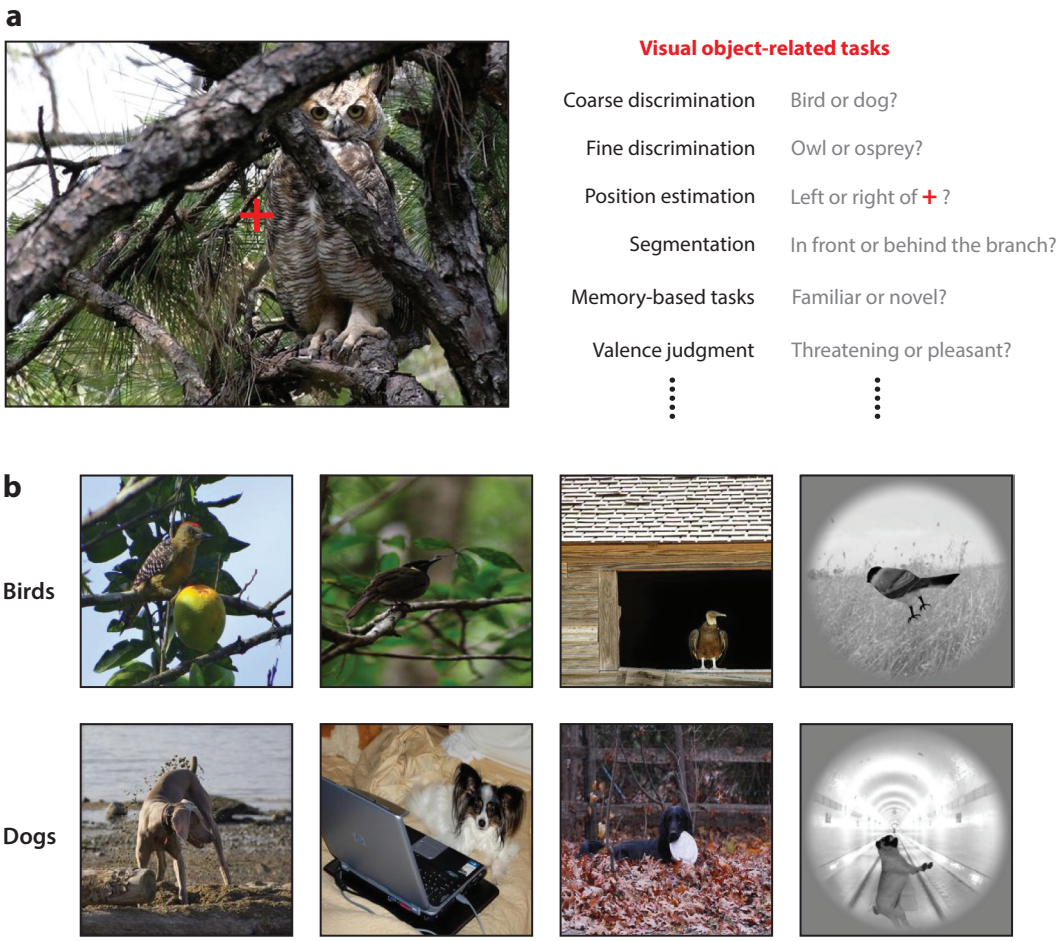

Figure 1

Probing visual object perception through diverse behavioral tasks. (*a*) (*Left*) An image—a two-dimensional pixel grid of RGB luminances—comparable to a photoreceptor-transduced spatial pattern of physical (photon) energy striking a person's retinae just after they turn their head to look up. (*Right*) A series of visual object–related perceptual report tasks, highlighting the multifaceted ways to investigate a subject's perceptual inferences about the latent content of the world from the image alone. These tasks span from coarse-category distinctions (such as discerning between the presence of a bird or a cat in the image) to more complex evaluations linked to memory and emotional responses (for instance, determining the familiarity of the bird or assessing its emotional valence). (*b*) Object recognition is algorithmically challenging because the same object category (i.e., the same type of latent cause) can generate a potentially infinite set of images, and successful behavior depends on inferring the presence of that object for any such image. Examples of images from the categories bird (*top row*) and dog (*bottom row*) are shown to demonstrate this challenge conceptually.

around objects. For instance, one could precue the subject about the types of objects to expect, and this could be done either explicitly (e.g., "You will see either a bird or a dog next") or implicitly (e.g., testing a block of many "bird versus dog" trials). Indeed, the effects of precueing have been extensively studied in the attention literature (Zhang et al. 2011). Human observers can also be asked to only report the object on a posttrial questionnaire or discrimination task. This article focuses on the postcueing, minimal-memory paradigm in which subjects enter each trial with many possible object categories to entertain (typically at least eight), and the question of which object was present is asked immediately after a test image. We consider this paradigm to put subjects

in a default attentional mode in which spatial attention (Maunsell 2015) is at the center of the scene (it is implicitly precued) and feature attention (Maunsell & Treue 2006) is also in a default mode in that the visual system can emphasize no single set of features due to the large number of potential objects and the associated complexity of features that must be handled to succeed in the task. We do not mean to imply that spatial and feature attention phenomena should not be part of a complete understanding of visual processing and visual intelligence, only that the mechanisms underlying those attentional phenomena are in reasonably natural default modes for most of the empirical studies that we discuss below.

As discussed above, core object recognition focuses specifically on the 100–200-ms viewing duration timescale, and thus, we review the mechanisms that are most relevant for that timescale. Longer viewing of images and videos will likely require additional mechanistic components beyond core recognition, including mechanisms for directing eye movements and integrating the information from each sampled image. Beyond object category and identity, other object-related latent variables such as object size, position, rotation, color, and material properties not only affect human estimates of object identity (for discussion, see Bracci & Op de Beeck 2023), but are themselves variables of objects that humans must also often accurately infer and are within the scope of core object recognition. In addition, the values of other object-related latent variables, such as object motion trajectory and velocity, could impact object-identity estimates and are also within the scope of core recognition.

In sum, a primary output of core object recognition is the contents of the subject's perceptual state causally induced by each image (e.g., the values of the set of object-related latent variables above). Key operational measures of this output include the subject's behavioral reports of those contents given a task paradigm (i.e., a way to trigger such reports). That is, we say that each image causes a particular perceptual state and its associated behavioral reports. For example, the presentation of an image of a cat will reliably produce the behavioral report of "cat," and removal of that image (e.g., presenting a full-field gray image) will reliably eliminate that behavioral report.

Our review is primarily focused on progress in understanding the primate brain mechanisms that underlie core object recognition. Given the above definitions and paradigms, it should be clear that core object recognition is not the entirety of what one might want to call object recognition, and it is certainly not all of visual perception. Nevertheless, the progress outlined below suggests that, somewhat fortuitously, a very large fraction of human ability to estimate the values of object latent variables (see above) and the visual processing that underlies many tasks beyond object recognition can be understood via machine-executable models that come out of this approach of solving core object recognition first.

1.2. What Do We Mean by an Understanding of Core Object Recognition?

Much of scientific understanding is in the form of reproducible models (Kuhn 1962, Popper 1934), ideally coupled to robust theoretical frameworks. Thus, any understanding of core object recognition should minimally include models that can potentially explain and predict empirical patterns of behavior for any image in the core recognition input domain (central 10°, <200 ms). The field does not agree on all model desiderata (e.g., compactness, explainability to others). Thus, the field does not fully agree on what comprises an understanding. In this review, we focus on models with three primary desiderata: (*a*) high reproducibility (i.e., models that, for any image, produce the same predictions in the hands of other scientists), (*b*) high empirical accuracy at the behavioral level (i.e., models whose predictions on new images tend to match the empirical observations of behavior, e.g., to match the pattern of successes and failures over images, where success is defined with respect to the ground-truth objects that generated the test images), and (*c*) brain-mapped mechanisms (at a particular level of resolution, defined below) with high empirical accuracy at the

neural level (i.e., the model predictions tend to match the empirical observations at the mapped level of resolution).

Sensory-computable model: an image-computable model that can be generalized to any sensory system

Image-computable model: a machine-executable system that can take any image as input and produce neural and/or behavioral predictions as outputs

We note that a model does not need to meet all three desiderata to be useful. For example, models that meet desiderata *a* and *b* would contribute to cognitive science. Models that meet desiderata *a* and *c* would contribute to neuroscience. However, models of the integrated set of neural mechanisms that underlie core object recognition must ultimately meet all three desiderata. To meet desideratum *a*, we focus on machine-executable (also known as computable) models that define a precise procedure (usually specified in software) that can be readily shared with other scientists to produce the same model predictions in different laboratories. As such, computable models, as defined in this review, have very high reproducibility.

In core visual object recognition (a behavioral capability), computable models must minimally take images (i.e., spatial patterns of photons) as input and produce behavioral reports in response to each image as output. Models that can make predictions (e.g., behavioral report predictions) for any given image are referred to as image-computable or, equivalently, sensory-computable models (see the sidebar titled SMART Models). Image-computable models are scientifically crucial because they engage the full complexity of all images (including all natural images), and they are precisely reproducible in the hands of other scientists and are thus independently testable (Yamins & DiCarlo 2016).

In sum, any sensory-computable model that accurately predicts the primate patterns of core object recognition behavior would, to us, constitute a potential causal scientific understanding of core object recognition. We note that some view this as necessary but not sufficient for understanding. We are not opposed to that view, but we strongly oppose the view that such models are not even necessary (for a discussion of this issue, see Schrimpf et al. 2020).

We do not mean to imply that only one such model exists (an infinite number exists). Nor do we mean to imply that other model desiderata are not potentially useful. Indeed, we are particularly

SMART MODELS

Sensory Computable

Predictions can be computed for any sensory input. For SMART models of core visual recognition, that sensory input is the spatiotemporal pattern of photons on the central 10° of the retinae. The model should include at least one behavioral report paradigm, like subjects' reports of object-associated latent variable values such as category, position, and pose.

Mechanistic and Anatomically Referenced

All major model components are mapped (i.e., permanently assigned) to a part of the brain. For ventral stream SMART models, the primary brain areas of interest are the four cortical areas of the ventral stream (V1, V2, V4, and IT), along with the retina and lateral geniculate nucleus. Current mappings are limited to the type II, and not the type III, level of mechanisms, which treat each layer of the models as a collection of neurons from a specific brain area without specifying any further level of detail about their connectivity with each other or to other brain areas (see Section 1.3).

Testable

The model will make falsifiable predictions about empirically measurable neural activity and behavior corresponding to any given test image. Successful predictions will support our field's belief in a particular model or set of models, and failed predictions will reduce that belief.

interested in models that are not only capable of explaining the primate behavioral pattern resulting from any sensory input but also capable of explaining how different parts of the brain work together (i.e., the underlying neural mechanism) to produce those behavioral patterns at various levels of detail. Next, we elaborate on what we mean by mechanistic models of object recognition.

1.3. What Is a Mechanistic Understanding of Core Object Recognition?

Above, we first explain what we mean by core object recognition—which is operationally defined as a sensory input domain (~10°, <200 ms) and a set of behavioral capabilities within that domain (DiCarlo et al. 2012). We then emphasize that a scientifically tractable understanding of core object recognition must centrally include reproducible, image-computable models that accurately explain and predict the patterns of core object recognition behavior and the neural mechanisms underlying those behavioral capabilities. However, it is not immediately clear what should comprise a neurally mechanistic understanding of that set of capabilities. Indeed, one can study the mechanisms of any behavioral capability at many different underlying levels (Churchland & Sejnowski 1988), and the literature demonstrates myriad observations about neurons and their connections that are likely related to the mechanisms of object recognition. This includes, for example, many reports of interesting neural functional phenomena associated with visual processing. A partial taxonomy of such reports includes spatial receptive field (RF) phenomena of visuocortical neurons (Rust et al. 2005), surround suppression phenomena (Jones et al. 2001), repetition suppression phenomena (Miller et al. 1991), various stimulus selectivity phenomena in neuron responses in key visual processing areas (Gallant et al. 1996, Levitt et al. 1994, Logothetis et al. 1995, Pasupathy & Connor 1999, Tanaka 1996, Tsao et al. 2006), and many other such seminal discoveries, far too numerous to list here.

Given this wealth of prior work, neurally mechanistic computational models are essential to integrating these myriad phenomena into a simulation of the system—from the sensory image, to multiple interacting neuronal subsystems, to behavior. But what is a sufficiently mechanistic model? Suppose that one was to deliver a computable algorithm (a type I system) that could take in any retinal image along with a prompt of a task goal, and it was empirically demonstrated that the behavioral output of that model in response to each input image precisely matched—that is, could precisely and accurately predict—human perceptual report for any new image. Would the source code of that system count as a satisfactory explanation of the mechanisms? We guess that, for most neuroscientists—ourselves included—this would not be a satisfactory mechanistic explanation.

Now suppose that a similar algorithm was constructed to also have a set of internal modules that each empirically behaved like a network (Ungerleider et al. 1982) (**Figure 2**) of specific visual brain areas (e.g., areas V1, V2, V4, etc.) (**Figure 2**) that were anatomically mapped to the hierarchical organization (Felleman & Van Essen 1991) of the primates' visual cortex. For example, just like in the brain, the algorithm's V1 module was activated slightly before its V2 module, etc. Setting aside the question of its empirical accuracy, this type II system is now at least slightly engaged in the question of mechanistic explanation.

Going further, now suppose that a type II model of visual areas was constructed to consist of only approximations of individual simulated neurons in each of those areas and their connections with other model neurons in the other areas (a type III model). That is, the new overall model would be a collection of model neurons, organized in a collection of model visual regions, that work together to give rise to a computational simulation of how those neurons process any image to give rise to a behavioral report. Clearly, this type III model is strongly engaged in the question of mechanistic explanation. Unlike the type I system (see above), this system is not only an algorithm—it is also a model of the integrated set of mechanisms. It is a mechanistic scientific hypothesis.

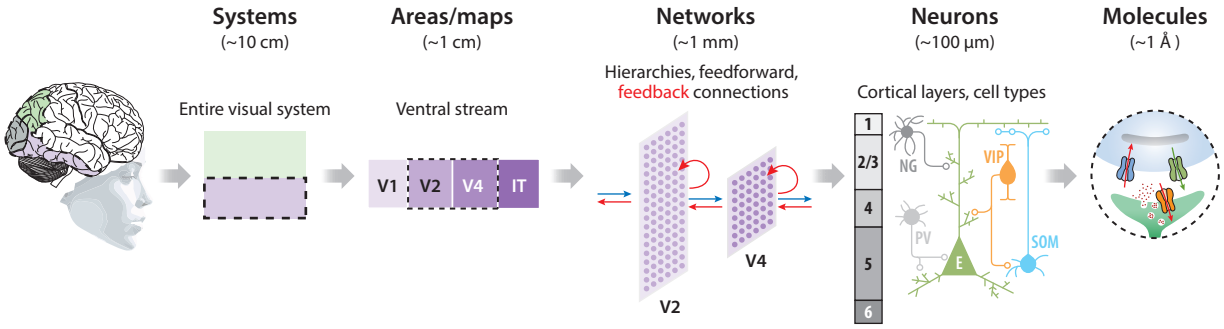

Figure 2

Constituents of a mechanistic understanding of object recognition with increasing levels of detail. The image shows a gradual progression from the systems level (the dorsal and ventral stream of the primate visual system shown here), to areas (ventral stream areas are shown), to networks (feedforward and recurrent connectivity within the extrastriate areas V2 and V4 are shown), to neurons (a schematic of arrangement within a cortical layer is shown) and molecules (a single synapse is shown). A thorough mechanistic understanding should gradually incorporate all levels [as suggested by Churchland & Sejnowski (1988)] of detail in a model. Abbreviations: E, excitatory neuron; IT, inferior temporal cortex; NG, neurogliaform cells; PV, parvalbumin-type inhibitory neuron; SOM, somatostatin-type inhibitory neuron; VIP, vasoactive intestinal peptide neuron. Image for neurons adapted from **http://knowingneurons.com**. Synapse image image adapted with permission from **https://scidraw.io/**.

Continuing even further, now suppose that a type III model was constructed to also incorporate detailed cortical layer–type structures and connectivity anatomy, different morphological and genetically defined neuronal cell types and associated synaptic transmission mechanisms, and biophysically verified dendritic models (**Figure 2**). This new type IV model would make quantitative contact with biophysics, thus linking to already-agreed-upon fundamental notions of a mechanism. In that sense, type IV models that successfully integrate all of these levels would, in effect, achieve a guiding dream of our field—to accurately and causally bridge from molecules to minds in visual object recognition. For example, an accurate type IV model would allow us to predict the precise changes in object perception that would and would not result from specific molecular interventions.

The overall point is that we are not pursuing any single set of mechanisms of object recognition. Instead, this biological capability—like all cognitive capabilities—can be explained at increasing levels of mechanistic detail. It is in this context that we next outline the state of our current mechanistic understanding, as captured in reproducible, sensory-computable models. We expect that our field will increasingly develop ever more precise models that make contact with ever-finer spatial scales (see Section 4). The current leading models (reviewed in Section 2) are type III explanations of mechanisms (see above).

2. SMART MODELS OF THE MECHANISMS OF CORE VISUAL OBJECT RECOGNITION

As outlined above, a critical rallying goal in understanding object recognition is the building of accurate models of the integrated set of underlying neural mechanisms and their support of object recognition behavior. This is an incredibly ambitious scientific goal: The expected generalization regime is effectively infinite—a successful hypothesis (i.e., model) must be accurate for any pattern of photons that impinges on the central 10° of the retinae, must accurately explain any object-related perceptual judgment that can be accomplished within 200 ms of viewing time (see Section 1.1), and must ultimately explain all of the functionally relevant neural phenomena in that

same spatial and temporal window—at least at the specified level of mechanistic resolution (see Section 1.3).

Because the term model is used in many ways, we aim to be more precise in this review. In particular, we seek models that are sensory-computable, mechanistic, anatomically referenced, and testable (referred to as SMART models; see the sidebar titled SMART Models). With this perspective, the goodness of our understanding of core object recognition (equivalently, the goodness of our current leading SMART models) should and can be primarily gauged by the accuracy with which these models explain and predict the myriad existing and future findings from all the relevant underlying brain components in the very broad regime outlined above. In this section, we summarize where the current leading SMART models of core object recognition came from and the neural and behavioral observations that they have been shown to explain and predict. In Section 3, we summarize explanatory gaps that still need to be bridged with new SMART models. In Section 4, we outline strategies to develop the next generations of SMART models.

SMART models: Sensory-computable, Mechanistic, Anatomically Referenced, and Testable models, which may be built by neuroscientists or inherited from AI system builders and then modified and mapped to the brain

2.1. A Sea Change in Neuroscience's Approach to Understanding the Mechanisms of Object Recognition

Many neuroscientists have been trained in bottom-up approaches where it is assumed that the study of low-level anatomical building blocks of a brain system (synapses, neurons, connectivity patterns, etc.) and the study of simplified functional phenomena (tuning functions, parameterized stimuli) can ultimately be pieced together to derive a type IV mechanistic model of core object recognition. As we describe below, that approach has now been reformed—top-down integrated models that aim to achieve capabilities like object recognition are now providing the scaffold to explain and understand the myriad bottom-up measurements.

Importantly, however, some bottom-up work in primates set the foundation for that sea change. In particular, several decades of neuroanatomical cortico-cortical tracing studies (Felleman & Van Essen 1991), neuronal lesion studies (Phillips et al. 1988), and neural recording studies identified the set of cortical processing stages collectively referred to as the ventral visual stream (Gross et al. 1972, Hung et al. 2005, Logothetis et al. 1995, Majaj et al. 2015, Tanaka 1996) as critical for core object recognition. The ventral stream consists of the primary visual cortical area V1, area V2, area V4, and the IT cortex (**Figure 2**). The input to this ventral stream starts at the retina, followed by further processing at the lateral geniculate nucleus of the thalamus, which then projects predominantly to cortical area V1, the first stage of the ventral stream.

Exploration into the nature of neural representation (i.e., the population pattern of neural firing in response to an image) in each of these cortical areas started with the seminal findings from Hubel & Wiesel (1962, 1968) in the cat primary visual cortex and macaques (Hubel & Wiesel 1968) and has since extended up to the apex of the ventral stream (Hung et al. 2005, Logothetis et al. 1995, Tanaka 1996, Tsao et al. 2006). Several organizing observations have been repeatedly made in the ventral visual pathway. For instance, researchers have observed an increase in the RF size of neurons along the hierarchy and a corresponding delay in mean neuronal response latency. In particular, in the central 10°, RF sizes progress from ~1° (V1), to ~2° (V2), to ~4° (V4), to ~10° (IT). Latencies progress from ~50 ms, to ~60 ms, to ~70 ms, to ~90 ms, respectively (DiCarlo et al. 2012; Gattass et al. 1981, 1988; Op De Beeck & Vogels 2000). In addition, the stimulus selectivity (i.e., how narrowly tuned to a specific type of natural stimuli or stimulus features neurons are) also tends to vary across these pathways. Specifically, while V1 neurons have small RFs and are nearly optimally driven by oriented (Gabor) patterns of light (Ringach et al. 2002), V2 neurons show preferential activations for various textures (Freeman et al. 2013), V4 neurons for curvatures (Pasupathy & Connor 1999), and IT neurons for a range of semantically meaningful concepts like

faces (Tsao et al. 2006) and bodies (Vogels 2022). Most of these observations were conducted with a limited set of hand-crafted, parametric images. In large sets of natural images, IT neurons have much more heterogeneous stimulus selectivity (Hung et al. 2005, Majaj et al. 2015). Implicit in many of these studies in areas V1, V2, and V4 was the observation that the selectivity properties at each stage of visual processing were approximately spatial shift invariant—that is, different neurons had the same functional selectivity as others (e.g., a preference for rightward-tilted Gabors) but operating in parallel at a fully tiled set of locations across the visual field.

Artificial neural network (ANN): a machine-executable system made up only of connected sets of weighted summation nodes (neurons)

Together, these bottom-up observations, along with anatomical tracing studies, pointed to a stacked, feedforward architecture with complete sets of shift-invariant neural spatial filters at each cortical stage as the scaffold of ventral visual processing (reviewed in DiCarlo et al. 2012). That scaffold architecture is today known as a deep convolutional neural network (DCNN), a particular subtype of artificial neural network (ANN) models (Yamins & DiCarlo 2016). Historically, the DCNN architectural family of models descended from work as far back as Fukushima (1980) and later work by LeCun & Bengio (1995), Riesenhuber & Poggio (1999), and Rumelhart et al. (1986). Nevertheless, despite 40 years of such bottom-up work following the seminal work of Hubel & Wiesel (1962, 1968), the field had not produced models—DCNNs or otherwise—that could solve the hard problem of core object recognition.

However, beginning in 2012, the field of visual neuroscience witnessed a sea change in approach. This change began with the emergence of some ANNs that began to rival primates in object categorization tasks. These ANNs were architecturally inspired by the ventral stream in that they were all DCNN subtypes of ANNs, thereby incorporating evidence from the bottom-up approach, as outlined above. Importantly, however, these new, high-performing DCNN models were also guided by a top-down behavior capability goal—successful assignment of each image to one of many object categories (e.g., Russakovsky et al. 2015). Progress toward that goal was fueled by optimization techniques that allowed the setting of the myriad network parameters that the bottom-up neuroscience functional phenomena could not determine (Krizhevsky et al. 2012; Yamins et al. 2013, 2014). These (DCNN) ANNs turned out to have unprecedented high performance on object recognition tasks and can be considered a key breakthrough point in the evolution of SMART models. The advent of high-performing DCNNs traces back not to a single event but rather to a combination of improvements around labeled image data availability, compute availability, architectural modifications, and optimization improvements. For a more comprehensive history of those developments, we point the reader to Yamins & DiCarlo (2016).

The key advances with respect to SMART models of primate core object recognition were demonstrated between 2012 and 2014. First, due to their demonstrated behavioral-level successes, some DCNNs were quickly elevated to the current best explanations of object recognition at the mechanistic model type I level (see Section 1.3). Second, the arrival and availability of these high-performing DCNN models enabled researchers to discover—surprisingly to many—that some of these models were among the leading hypotheses at mechanistic type II and type III levels (Section 1.3). Most notably, it was discovered that the internal simulated neurons in these models were highly functionally similar to biological neurons along the ventral visual stream and significantly better than previous models in the field (for reviews, see Schrimpf et al. 2018, Yamins & DiCarlo 2016).

Next, we focus on the successes and current weaknesses of this top-down, achieve-behavioral-capability-first approach (also known as the performance optimization approach) (Yamins & DiCarlo 2016) to building a mechanistic understanding (i.e., SMART models). We see this approach as important and synergistic with the more traditional bottom-up neuroscience approaches (as described in Sections 4 and 5).

2.2. Empirical Tests of Current SMART Models

SMART models of object recognition have made significant strides toward emulating human object recognition capabilities. For an up-to-date, complete list of the currently leading SMART models and their empirical evaluation, we point the reader to the open science Brain-Score platform [**http://brain-score.org** (Schrimpf et al. 2018)] (note that this platform refers to SMART models as brain models).

2.2.1. Behavioral prediction tests of SMART models. Initially, as outlined above (Section 2.1), the foremost objective of the ANN precursors of SMART models was to achieve human-level performance in terms of mean behavioral accuracy over many categories, which has been a primary benchmark in computer vision in assessing the efficacy of these models. Remarkably, some ANNs have not only reached but, in some instances, surpassed the threshold of mean primate accuracy (Dosovitskiy et al. 2020), at least for situations that are not substantially different from what is typical (but see Barbu et al. 2019).

One can easily imagine a computer vision system that matches or exceeds mean primate performance but that makes mistakes that are not primate-like (e.g., think of the barcode reader at the supermarket checkout). In contrast, a fully accurate SMART model must, by definition, not just match overall mean primate performance but also make the same mistakes that humans make. Note that this is where the neuroscience or cognitive science definition of an accurate model (empirical alignment with the brain and its output) differs from the computer vision definition of accuracy (performance relative to ground truth). In this regard, it is highly nontrivial that some of the high-performant DCNNs (in the computer vision sense) also turned out to have unprecedented good alignment with independently measured patterns of human object recognition behavior. For example, for some ANN systems, objects that are difficult to discriminate are also difficult for humans to discriminate, and objects that are easy to discriminate are also easy for humans to discriminate. Studies using careful quantitative testing report that some DCNN models are statistically indistinguishable from humans and monkeys (see human and monkey comparisons in **Figure 3**) at this level of behavioral comparison (referred to as the consistency of object-level confusions) (Rajalingham et al. 2015) (see **Figure 4*b***). Indeed, such strong empirical alignment observations are part of what elevates some—but not all—ANN systems from just being brain-inspired technology drivers to being scientific SMART models of primate core object recognition.

The current leading SMART models and humans make surprisingly comparable errors (on object categories and individual images), suggesting a deeper, structural similarity in the way visual information is processed. This behavioral-level alignment extends to more nuanced aspects of visual recognition and hierarchical processing of visual input, further underscoring the parallels between artificial and human visual cognition (Jacob et al. 2021). However, even the leading SMART models are not entirely behaviorally aligned with primates in all respects (see Section 3).

2.2.2. Neural response prediction tests of SMART models. One quantitative way to determine whether the neural mechanisms inside a candidate SMART model explain those at work in the ventral stream is to measure the functional similarity of neural representations in both of those systems. Such comparisons can be done in several ways (see the sidebar titled Testing the Neural Alignment of SMART Models), and methods and statistics around such comparisons are an active area of research (Kriegeskorte et al. 2008; Schrimpf et al. 2018, 2020; Yamins et al. 2014). At their core, all of these empirical tests ask about the ability of the simulated neuronal population in SMART model area X (e.g., SMART model area V4) to predict the neuronal population in that same ventral stream area. The notion of prediction here refers to the testing of images that were

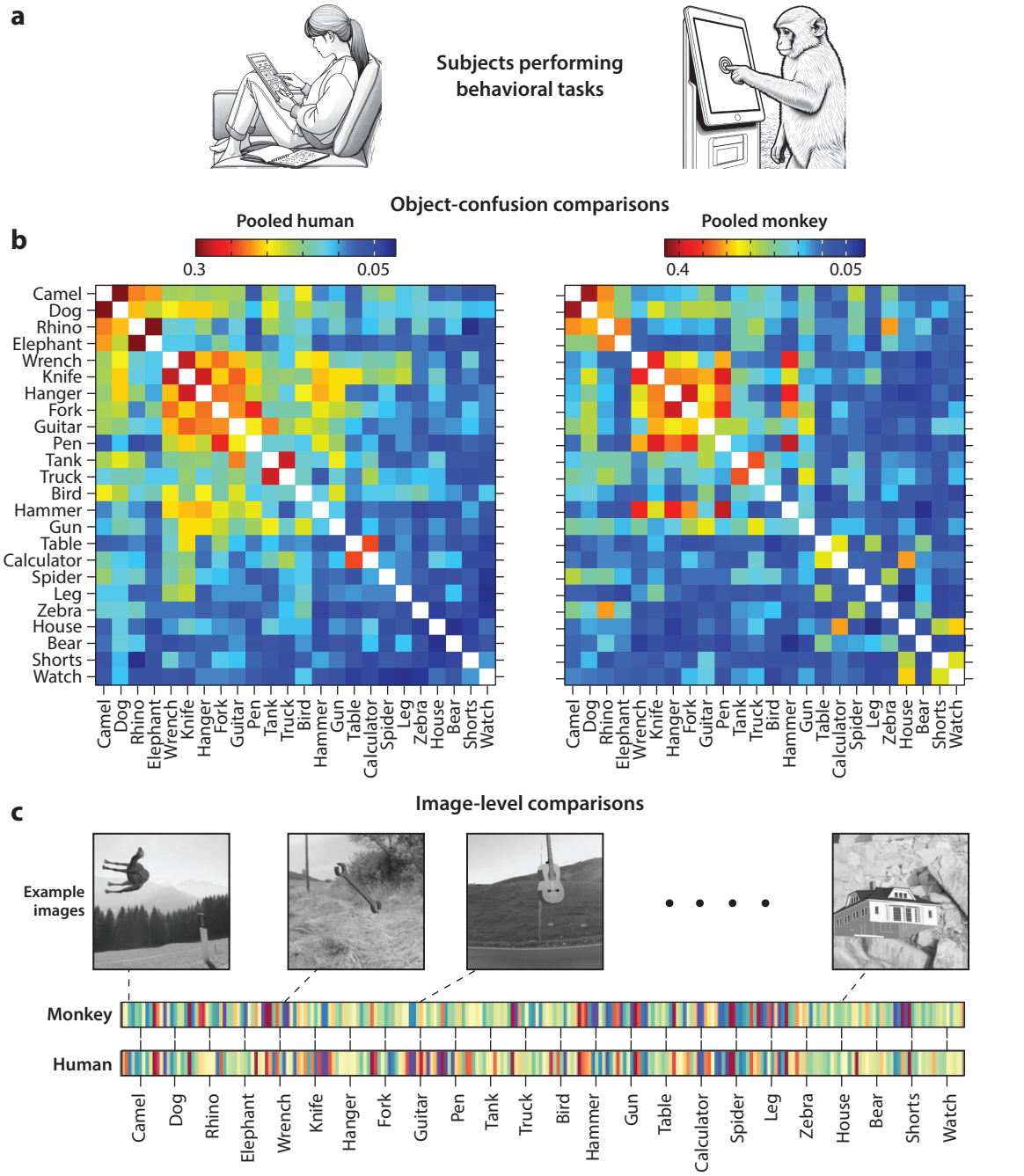

Figure 3

Comparison of object recognition behavior between monkeys and humans. (*a*) Schematic of two subjects, a human and a rhesus macaque, participating in a behavioral task. (*b*) Comparison of object-confusion patterns across pooled human (*left*) and monkey (*right*) populations (for details, see Rajalingham et al. 2015). (*c*) A finer-grain comparison can be performed at the image level (example images shown). Each value is approximately equivalent to the behavioral accuracy of determining, among a set of possible objects, which object generated that particular image (for details, see Rajalingham et al. 2018).

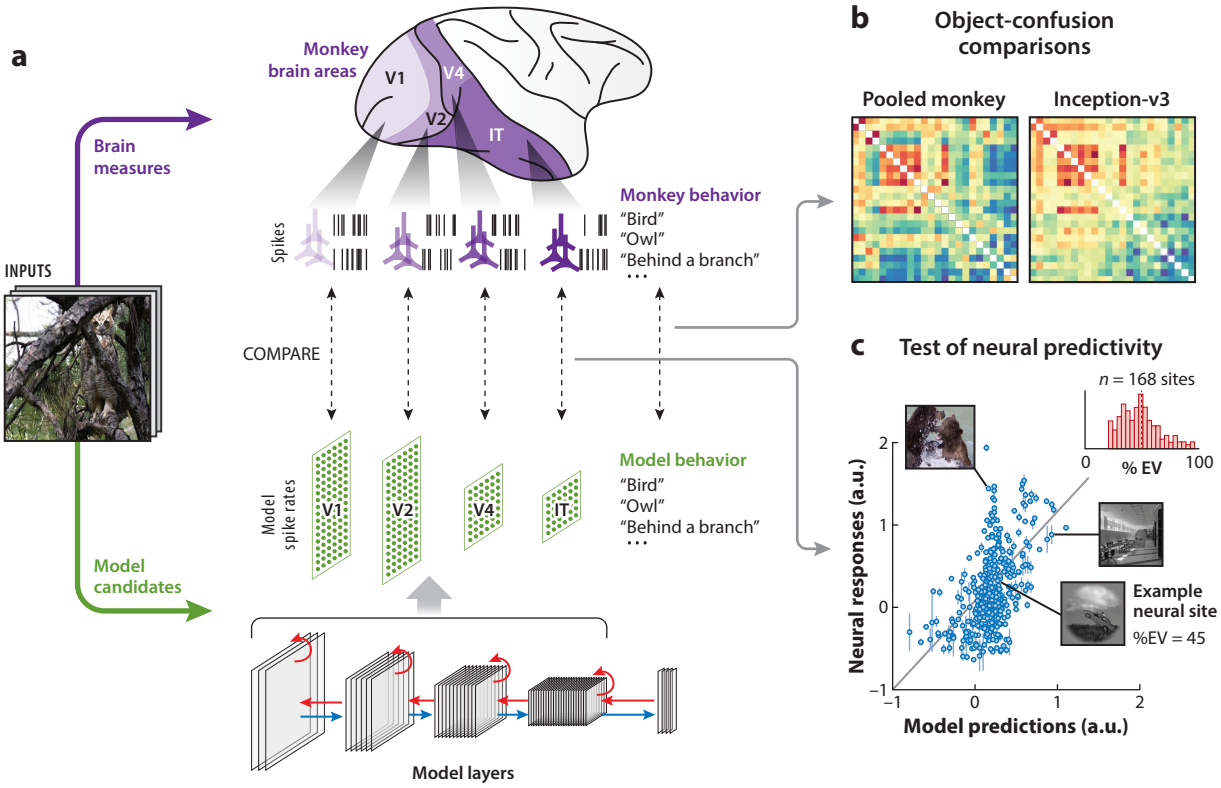

Figure 4

Evaluating the alignment of SMART models with primate behavior and neural responses. (*a*) Current SMART models (see the sidebar titled SMART Models) are derived from brain-mapped, image-computable deep CNNs. The predictions of SMART models—for a proposed experiment—are obtained simply by performing the same experiment on the model: for instance, by presenting a planned set of images as the input to the model and recording the responses of individual model neurons from a model brain area (e.g., IT) or by recording its behavioral responses. To assess the empirical alignment with biology, these predictions are compared with the results of that experiment, scored with a quantitative metric. Each alignment test is referred to as a benchmark. When multiple benchmarks are performed on one SMART model, this is referred to as integrative benchmarking (Schrimpf et al. 2018, 2020). Here, we illustrate just one behavioral comparison (*b*) and just one neural comparison (*c*). (*b*) Comparison of the monkey behavioral object-level confusion patterns and an inception-v3-derived SMART model (Szegedy et al. 2016). The metric of alignment here is the correlation over all of the values in the two matrices (see Rajalingham et al. 2018). (*c*) An image-level neural response predictivity test (here, for one study of the IT cortex). The scatter shows results from one example IT neural site and the SMART model–predicted responses of this site (to do this, the model must be mapped at the level of single units; see the sidebar titled Testing the Neural Alignment of SMART Models for how that is done). Each dot is the model-predicted response (*x* axis) versus the actual neural response (*y* axis; mean rate in a time window, averaged over repetitions of that image). The elemental alignment metric here is the fraction of image-response variance that is accurately predicted (EV), corrected for irreducible noise. The overall alignment metric is the median EV over all recorded neural sites in the dataset (*inset histogram*) (Yamins et al. 2014).Abbreviations: CNN, convolutional neural network; EV, explained variance; IT, inferior temporal; SMART, sensory-computable, mechanistic, anatomically referenced, and testable.

never used to estimate any of the SMART model's internal parameters and were never used to estimate the model-to-brain mapping parameters [that is, testing which simulated neuron(s) in the model correspond to the biological neuron(s) of interest].

Before describing some of those results, we note that, unlike SMART models, ANN or DCNN vision systems that do not have mapping commitments to brain areas cannot be tested in this way. This lack of commitment does not reduce the potential utility of these systems in other

TESTING THE NEURAL ALIGNMENT OF SMART MODELS

Neural Response Measurements

Model. Measurements of SMART model neurons are made by presenting test images and recording activation values of neurons in the model area—known as extracting features from a specific layer in artificial intelligence.

Brain. Experimental recordings of individual neural sites in a brain area are made using the same test images. Spikes are counted in a latency-adjusted time window (typically 70–170 ms post–image onset), averaging over repeat presentations. SMART models have also been compared at finer temporal resolution (Kar et al. 2019).

Mapping SMART Model Neurons to Biological Neural Units

Current SMART models are anatomically referenced at the brain area level (see the sidebar titled SMART Models). To make finer-spatial-grain predictions, SMART model neurons must be mapped to biological neurons. There are several methods to do this, each with pros and cons (Arend et al. 2018, Kar et al. 2019, Klindt et al. 2017, Yamins et al. 2013). However, all mapping methods assume a linear relationship between model neurons and biological neurons. Once determined, the mapping is frozen for evaluation.

Metrics to Assess the Goodness of Model Predictions

Neural predictivity measures how well predicted neural responses match actual measured responses, typically in units of explained variance, R2, corrected for nonreproducible variance.

Representational similarity analysis, pioneered by Kriegeskorte et al. (2008) (see Nili et al. 2014), compares distance matrices constructed from neural measurements and corresponding SMART model neural population responses.

Centered kernel alignment was proposed by Kornblith et al. (2019) as a similarity measure invariant to rotation and isotropic scaling, but not all linear transformations, to compare model and brain population representations.

venues. Instead, it simply reflects a lack of engagement on the question of neural mechanism (see Section 1.3).

Beginning in 2013, it was discovered that the responses of SMART models' internal components—artificial neurons within each of the model areas (i.e., model layers)—often strongly align with the responses of their biologically mapped counterparts (Cadieu et al. 2014; Yamins et al. 2013, 2014). These and many later studies showed that current SMART models can predict ~50% of the explainable neural response variance. This was significantly better than models from a decade ago (Riesenhuber & Poggio 1999, Serre & Riesenhuber 2004) but still less than perfect (performing below the noise ceiling as estimated per neuron). It is important to note that how well a model neuron should predict an IT neuron recorded from a randomly sampled monkey depends on many prior assumptions (e.g., whether we are building a model of that specific monkey or an archetypal monkey) and is a matter of ongoing research.

Over the past decade, many studies in the ventral stream have either explicitly or implicitly replicated this core neural finding. For example, at the spiking neural level, recent SMART models predicted V1 responses to natural images with unprecedented accuracy (Cadena et al. 2019, Dapello et al. 2020), predicted specific types of shape tuning in V4 neural responses (Pospisil et al. 2018), and were reported to be the best predictor of anterior IT face-patch response (anterior medial) (Chang et al. 2021). Other studies have used SMART models to predict functional aspects of ventral stream neural representations as assessed by functional magnetic resonance imaging

(fMRI) (Güçlü & van Gerven 2015, Khaligh-Razavi & Kriegeskorte 2014, Ratan Murty et al. 2021), electrocorticography (Grossman et al. 2019), and magnetoencephalography (Cichy et al. 2016).

Despite the diversity of stimuli and methods, it is challenging to tell if there is a trend for some areas of the ventral stream to be better explained than others. This is compounded by the fact that different areas have different functional dimensionality (in the models and likely in the biology as well), which makes such comparisons dependent on the metrics used. To our knowledge, the best summary of the current state of SMART models of the ventral stream and its supported behavior is tracked on the open science Brain-Score platform (**http://brain-score.org**). While far from flawless, this platform is better than no tracking at all, and it continues to improve in its functionality and number of neural and behavioral benchmarks. Inspection of the Brain-Score benchmarks suggests that leading SMART models currently capture a large amount of neural functional response variance but that no model is yet fully accurate, even among the limited set of neural benchmarks that are available.

Taken together, what all of these studies imply is that the image-driven functional response profiles of individual neurons along the ventral stream are surprisingly similar to the functional profiles of their individual digital twin SMART model neurons. The representational tests imply that the population distributions of the different functional types of neurons are approximately matched. Indeed, the leading SMART models of the ventral stream are referred to as the leading models in part because they do very well on these neural functional comparisons—far better than earlier models.

2.2.3. Neural control tests of SMART models. "All models are wrong, but some are useful," an aphorism attributed to the statistician George Box, also applies to models of object recognition. Recently, the value of SMART models has been augmented by the goal-directed stimulus synthesis of images. For instance, Bashivan et al. (2019) demonstrated (**Figure 5*a***) that, by using a SMART model that included visual area V4, they could generate synthetic stimuli that drive specific, experimenter-chosen V4 neurons to response levels beyond what could be achieved by the previously known preferred stimuli for the region. They also showed that this approach could be used to target entire subpopulations of recorded neurons—demonstrating at least a partial ability to independently set each neuron at a desired activation state. These tests have been referred to as neural control tests because the goal is to drive or set (i.e., control) the neural activity level or levels to a particular, experimenter-chosen state. A critical observation from that study was the high correlation between the accuracy of the model predictions over naturalist images and the quality of neural control that could be obtained, suggesting that the neural prediction measures (Section 2.2.2) are reasonable proxy measures of the more applied goal of neural control. Related experiments have been carried out in other brain areas. For instance, Ponce et al. (2019) demonstrated that they could synthesize superstimuli for IT neurons that drive the activity of these cells beyond their usual response range (**Figure 5*b***). In fact, their results challenge the common terminology in the field, given that the superstimuli for a classical face-selective neuron do not resemble a typical primate face—paving the way for a new set of model-based intuitions for how to think about neuronal encoding spaces (but for divergent results in human fMRI, see Ratan Murty et al. 2021). Similar approaches have also been implemented in the rodent neuroscientific community (Walker et al. 2019).

2.3. Future Tests of SMART Models: Direct Neural Perturbations

Tests of SMART models should extend beyond behavioral and neural studies, using direct neural perturbations like optogenetics, electrical, and other interventions to understand the mechanistic

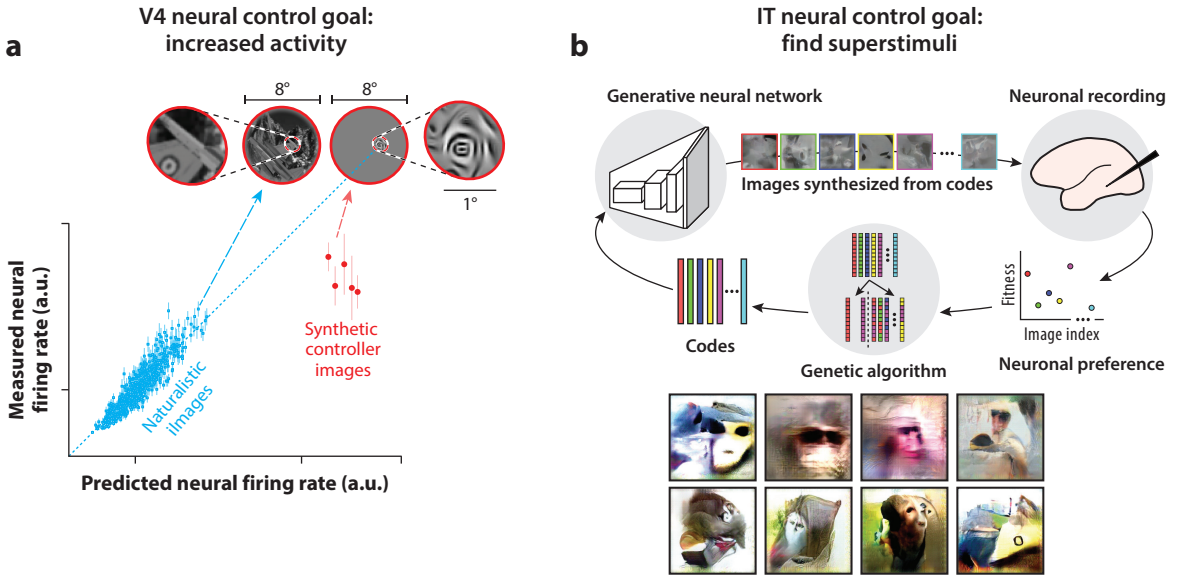

Figure 5

Examples of neural control in mid-level area V4 and in the inferior temporal (IT) cortex. (*a*) Model-guided generation of synthetic images was shown to increase the neural firing rate beyond the range observed in a large set of naturalistic images (for details, see Bashivan et al. 2019). (*b*) A schematic of the generative evolution method XDream (Xiao & Kreiman 2020), in which a deep generative adversarial network was used to synthesize images presented to monkeys. Neuronal responses were used to rank and optimize the image codes using a non-gradient-based optimization algorithm, here illustrated for a genetic algorithm. The bottom panels show examples of images evolved for face neurons (*top row*) and nonface neurons (*bottom row*) (for details, see Ponce et al. 2019).

role of brain components. Despite their potential, these tools currently offer limited and arbitrary levels of direct neural control (Jazayeri & Afraz 2017, Wolff & Ölveczky 2018), often reinforcing established conceptual causal models rather than distinguishing among alternative models. Yet the similarity between brain tissue and SMART models presents a unique opportunity to use these techniques to differentiate among and evaluate alternative SMART models. Advanced ANNs could simulate in vivo perturbations, promising better understanding and improved visual prosthesis strategies. This approach could bridge gaps between traditional experiments and innovative model testing.

3. KNOWN MISALIGNMENTS BETWEEN BRAINS AND CURRENT SMART MODELS

Despite the surprising empirical successes of the current family of SMART models described in Section 2, there is still misalignment between these models and the primate brain at both the neural and behavioral levels. Next, we review some known misalignments of the current ANN-based SMART models with primate neurobehavioral data. Here, we only discuss examples of neural and behavioral functional phenomena that current SMART models do aim to predict, yet fail. In Section 4, we discuss how next-generation SMART models could aim to predict even finer-scale phenomena, with successes and failures yet to be determined.

3.1. Behavioral Prediction Failures

Recently optimized ANNs solve object recognition tasks at unprecedented mean accuracies (He et al. 2016). However, as of a few years ago, no ANN exhibited patterns of successes and errors

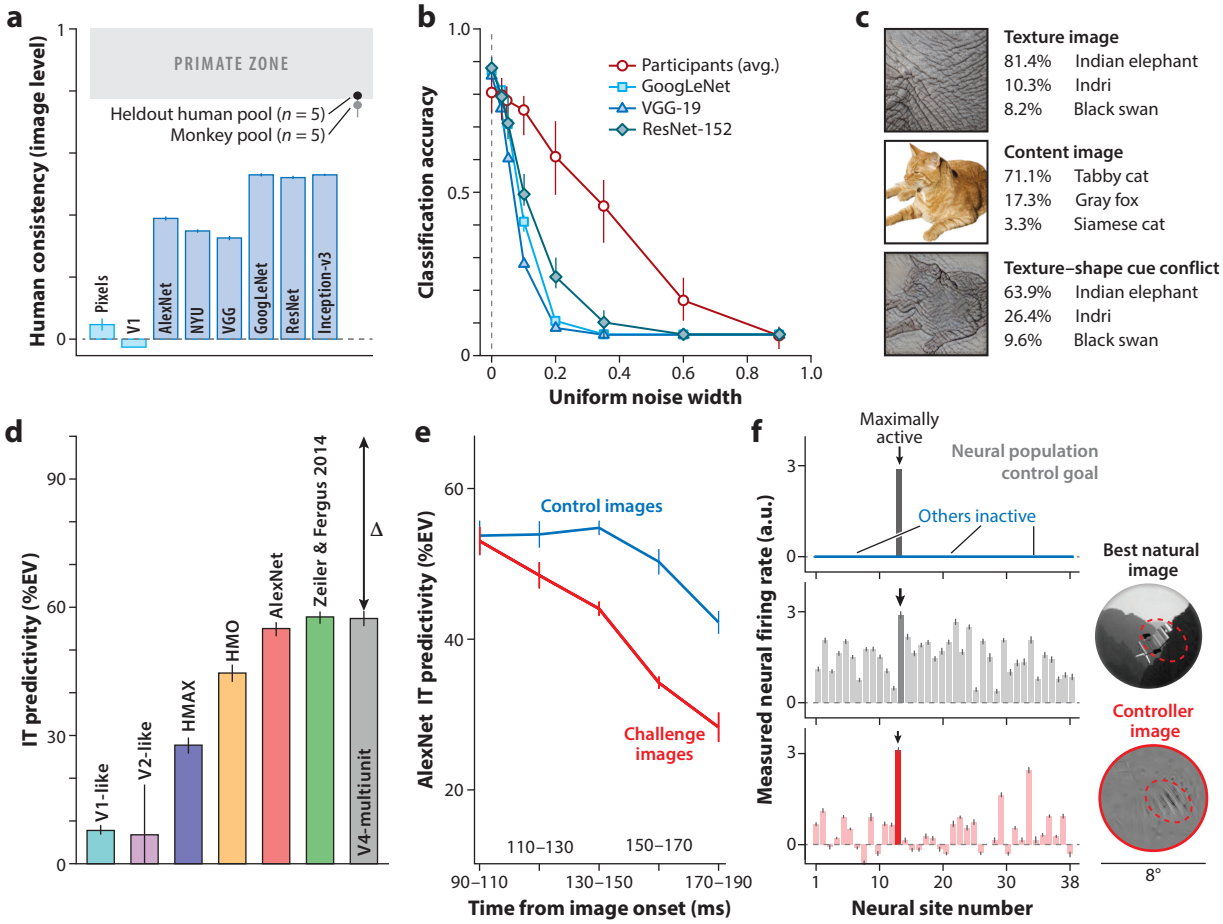

Figure 6

Explanatory gaps in behavioral and neural predictions. (*a*) The consistency of image-level accuracies during object discrimination tasks between humans and several tested models (see Rajalingham et al. 2018). (*b*) Changes in classification accuracy with increasing levels of image noise across VGG-19, GoogleLeNet, ResNet-152, and human participants demonstrate greater noise robustness in human vision (Geirhos et al. 2018a). (*c*) Classification of ResNet-50 of elephant skin (only texture cues), a normal image of a cat (consistent shape and texture), and an image with texture–shape cue conflict (shape of a cat, texture of an elephant) showing the so-called texture bias in leading SMART models at that time (Geirhos et al. 2018a). (*d*) Neural response predictivity in the IT cortex for SMART models at that time. The Δ denotes the explanatory gap (for the most up-to-date measures, check **http://brain-score.org**). (*e*) Relative to the early IT response (90–120 ms after stimulus onset), feedforward SMART models are poor at predicting the late IT population response (150–200 ms). This gap is particularly prominent for test images for which monkeys behaviorally outperform the models [challenge images (red line); AlexNet shown here] compared to images where models and humans have similar performance [control images (blue line)] (for details, see Kar et al. 2019). (*f*) The control objective of the OHP is to selectively increase the activity in one neural site while keeping responses of all recorded neural sites close to zero (*top row*). The middle row shows the naturalistic image that most closely accomplishes this objective. The bottom row shows a SMART model–driven synthetic controller image that performs much better but still does not fully achieve the OHP objective (Bashivan et al. 2019). Abbreviations: IT, inferior temporal; OHP, one-hot problem; SMART, sensory-computable, mechanistic, anatomically referenced, and testable.

across images that fully aligned (**Figure 6*a***) with human patterns measured over the same images (Rajalingham et al. 2018). More targeted looks into these misalignments have revealed specific shortcomings of ANNs that make them incomplete models of human behavior. We discuss the most prominent of these targeted analyses below.

First, Geirhos et al. (2018b) observed that some leading ANNs at that time (e.g., VGG-19, ResNet-152) were less robust (compared to humans) to the addition of Gaussian noise to images during object categorization (**Figure 6*b***). Interestingly, Geirhos et al. (2018a) also discovered that these ANNs relied more on the texture of the objects compared to their shapes (**Figure 6*c***), while humans typically rely more on object shape in comparable tests.

Second, is the behavioral susceptibility of ANNs to so-called adversarial attacks (Goodfellow et al. 2014). In brief, given the full observability of all ANNs, optimization methods have been used to search through high-dimensional pixel space and successfully find small-amplitude image perturbations that strongly change the behavioral output of the ANN (e.g., changing the output from "dog" to "church"). The (Euclidean) pixel amplitude of these attacks is typically less than a few percent of the distance between arbitrary natural images, and it was demonstrated that human behavior is largely [but not completely (Elsayed et al. 2018)] insensitive to the same changes. This suggests a potential mismatch of those early SMART models with human vision. At the time of this writing, tests on newer SMART models, which also have higher neural alignment (Guo et al. 2022, Schrimpf et al. 2018), have revealed that human perception can be surprisingly and strongly modified by similarly small-amplitude image perturbations (Gaziv et al. 2023). And, when properly compared, these current leading SMART models have far less behavioral misalignment with human perception than the original adversarial work (Gaziv et al. 2023). However, a gap nonetheless remains.

Third, other phenomena of visual perception are thought not to be well predicted by current SMART models. Examples include local versus global shape processing phenomena, the dependence of object classification on object part relationships, filling in illusory phenomena, and uncrowding phenomena (Bowers et al. 2022). However, many of these putative behavioral gaps have not been systematically tested. Scientific caution is warranted here, as SMART models continue to unexpectedly predict things that were not part of their explicit design and, thus, that one might not expect them to predict (Fan & Zeng 2023, Ngo et al. 2023). These are now active areas of model-to-human empirical comparison studies.

3.2. Neural Prediction Failures

As outlined in Section 2, current SMART models are surprisingly accurate at predicting neural responses in areas across the ventral stream, even at the single-neuron level. However, even the current best SMART models only predict 50–60% of the explainable variance in the neural responses in V4 and IT cortex (**Figure 6*d***) (for the most updated statistics, refer to Brain-Score). This exposes an apparent explanatory gap that still remains to be bridged. A more targeted investigation of these misalignments reveals that current SMART models are not fully accurate models of ventral visual processing, even at their currently intended mechanistic level (type III; see Sections 1.3 and 4). We discuss a few of these targeted analyses below.

First, when looking specifically into the functional subtypes of neurons, like face-selective neurons of the IT cortex, Chang et al. (2021) reported that CORnet-S (a leading SMART model) only predicts ∼50% of the explainable neural variance. In addition, the layers of the current SMART models and the brain areas of the primate ventral stream are not strictly hierarchically aligned, necessitating more careful investigation of how signals across these areas are integrated over time and how the models could explicitly implement these computations (Sexton & Love 2022). Second, the neural dynamics of current SMART models are clearly not in line with those of the ventral stream. For example, Kar et al. (2019), working at the spiking neural level, recently found that, while feedforward ANNs do quite well at predicting the IT population pattern in the early phase of the neural responses (90–120 ms after image onset), they are poor to moderate at

predicting the late phase of the neural population pattern (150–200 ms after image onset). As shown in **Figure 6*e***, this difference increases for images [labeled as challenge images by Kar et al. (2019)] where primates outperform baseline ANN models such as AlexNet. This observation is consistent with the lack of recurrent connections in these ANNs that other results suggest are critical in shaping the late phase of the IT response (Kar & DiCarlo 2021). Third, Bashivan et al. (2019) observed that current SMART models at that time did poorly at predicting V4 neural responses to strongly out-of-domain images, a finding also demonstrated for IT responses (Ponce et al. 2019). Bashivan et al. (2019) also observed that the one-hot population control paradigm (where the objective was to design images that only activate one neural site while not activating all of the other measured V4 sites) (shown in **Figure 6*f***, top), could not be perfectly executed (as shown in **Figure 6*f***, bottom).

These observations collectively point toward the inherent limitations of current SMART models, even at their currently intended mechanistic level, and emphasize the pressing need for iterative advancements in SMART modeling approaches to encompass the intricate nuances of the primate ventral stream neural machinery.

4. WHERE WILL THE NEXT GENERATION OF SMART MODELS COME FROM?

Over the past decade, significant progress has been made in developing SMART models of visual object recognition at the center of gaze. Importantly, AI engineering is helping to fuel that progress (see sidebar titled Role of Artificial Intelligence Engineering in Systems Neuroscience; **Figure 7**). However, these models have limitations. First, the current best models still do not account for (i.e., predict and control) 100% of the neural and behavioral functional measures of core ventral visual processing that they already aim to account for (see Section 3). Second, it is still unclear if simple variants of the current models can or cannot account for visual processing and visual behaviors beyond the central 10° and beyond what is achieved in the first ∼200 ms of visual processing in a default attentional state. Lastly, current SMART models do not yet map to—and thus cannot yet account for—the potentially different functions of different cortical layers, anatomical recurrences, and diverse neuronal cell types (including cells with different morphologies and genetic profiles).

ROLE OF ARTIFICIAL INTELLIGENCE ENGINEERING IN SYSTEMS NEUROSCIENCE

It is a striking observation that artificial intelligence (AI) engineering to performance optimize a ventral stream–inspired family of deep ANN models—but without further regard for the brain—returned a generation of neuroscientific models of the brain mechanisms that were more accurate than their predecessors (Cadieu et al. 2014, Yamins et al. 2014). Does this mean that neuroscientists should just sit back and wait for AI engineering to deliver the next generation of better neuroscientific models? In theory, it should be obvious that this trend must have limits—one cannot model all of biology without empirically studying biology. Indeed, while this remarkable upward trend continued for SMART models of the ventral stream from 2013 to 2016, we have already seen the turning point (**Figure 7**). Today, more accurate neuroscientific SMART models are deriving from a close collaboration between natural science experiments and AI engineering. However, other areas of systems neuroscience are, we believe, still on the upward trajectory in that even loosely brain-inspired AI engineering is indeed still producing the leading neuroscientific models (Kell et al. 2018, Schrimpf et al. 2021). We expect such trends to continue and then evolve in a similar way: An initial period of AI engineering–driven gains, followed by a period with a tight iterative loop between experiments and SMART model updates, is needed.

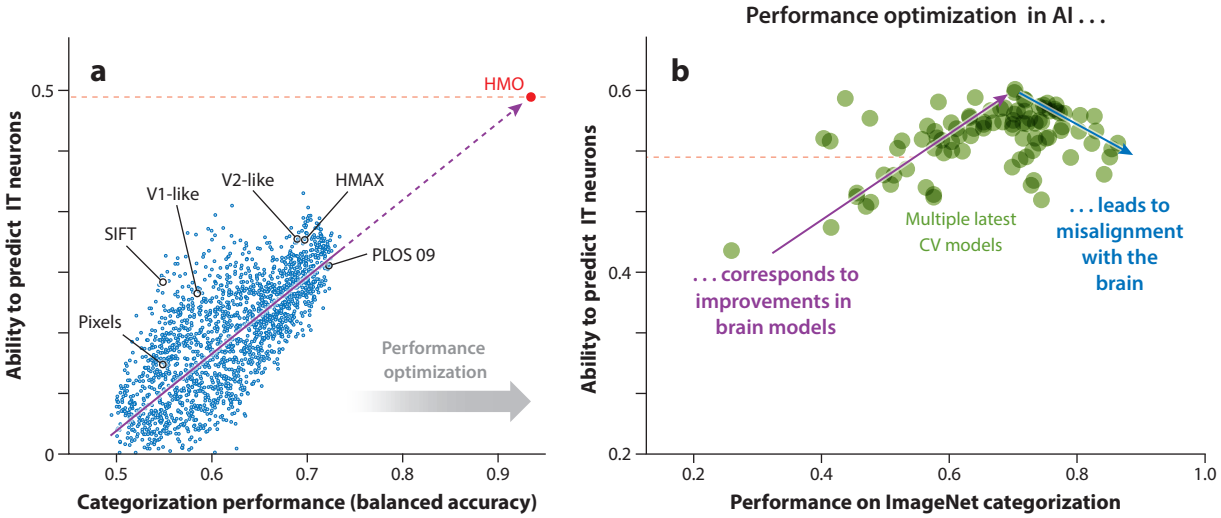

Figure 7

Relationship between object categorization performance and neural alignment of ANN models. (*a*) Each dot is a set of CNN model features. The *x* axis shows performance on an object categorization task (not ImageNet). The *y* axis shows IT neural predictivity of the model features. Noting this trend, shown in panel *a*, Yamins et al. (2014) used optimization methods to develop the HMO model and found that HMO's penultimate layer (*red dot*) explained unprecedented levels (∼50%) of IT response variance at that time. Panel adapted from Yamins et al. (2014). (*b*) Relationship between object categorization accuracy [ImageNet (Russakovsky et al. 2015) accuracy] and IT predictivity (see the sidebar titled Role of Artificial Intelligence Engineering in Systems Neuroscience; results taken from Brain-Score). Beyond HMO (panel *a*), improvements in the categorization performance of the overall model continued to produce a hidden feature layer that followed the trend that Yamins et al. (2014) had identified, leading to even higher IT predictivity (*purple line*). However, the trend did not empirically continue past 2017 (see the sidebar titled Role of Artificial Intelligence Engineering in Systems Neuroscience for perspective on this).Abbreviations: AI, artificial intelligence; ANN, artificial neural network; CNN, convolutional neural network; CV, computer vision; IT, inferior temporal.

We believe that these three challenges are tightly interrelated. For example, recall that model performance gains in core visual object recognition led to more accurate models of the neural mechanisms of the first 200 ms of visual processing (see Yamins et al. 2014, figure 1). Thus, we anticipate that the development of next-generation SMART models that achieve performance gains in other visual intelligence capabilities for which primates rely on longer timescales of visual input will lead to even more accurate predictions of longer-timescale neurobehavioral dynamic phenomena along the ventral stream.

Next, we highlight some ongoing and forward-looking activities and ideas in each of these three interrelated research directions.

4.1. Directly Improving Upon Current SMART Models of Ventral Processing

A basic recipe to search among alternative SMART models of ventral visual processing has been discussed by Yamins & DiCarlo (2016) and formalized in the deep learning framework by Richards et al. (2019). In brief, the four key components to be explored are the model architecture (the functional building blocks of the model), the behavioral objective [the capability goal(s) of the model, e.g., object categorization], the learning rules (how the model is optimized with its set architecture to accomplish the behavioral objective), and the ecological niche (i.e., the training diet used to try to achieve the behavioral objective).

Ongoing work aims to explore these components. The primary goal of many of these studies to date has not been to improve the empirical accuracy of current SMART models on adult functional

measures (Section 2). Instead, these studies (*a*) aim to find minimal conditions that might give rise to SMART models with similar empirical accuracy as ANNs discovered by computer vision via bio-implausible optimization methods and (*b*) extend the set of adult phenomena that SMART models accurately account for. This includes, for example, efforts to explore more plausible evolutionary selection mechanisms (Geiger et al. 2020), more biologically plausible learning rules that might unlock new hypotheses about postnatal visual development (Zhuang et al. 2021), more ecologically relevant experience histories (Barbu et al. 2019, Mehrer et al. 2021), and/or mechanisms that can also explain the topographic organization of the ventral stream (Dobs et al. 2022, Doshi & Konkle 2023, Lee et al. 2020, Margalit et al. 2023). These important normative activities each seek to develop new variants of SMART models that can explain not only how the ventral stream works the way that it does, but also why it works the way that it does and how it got to be that way.

In addition to this ongoing normative research, recent work has also been aimed at using adult functional data to directly guide the building of more accurate SMART models of these types of data. For example, Dapello et al. (2022) directly used neural recordings from the IT cortex to regularize the training of ANNs (alongside ImageNet categorization loss), leading to more human-aligned ANNs that better predicted neural responses on new monkey subjects, images, and behavioral patterns and also became more adversarially robust. In a similar approach applied to behavior, Fel et al. (2022) developed a neural harmonizer training method that aligns ANNs with human visual strategies while also enhancing categorization accuracy on new images. These direct model optimization approaches may not be sufficient on their own in the near term due to current data limits. Nevertheless, as experiments are beginning to produce ever-larger volumes of functional primate and human data, we suspect that this strategy will also be an important part of discovering next-generation SMART models of the ventral stream.

4.2. Evaluating Alternative SMART Models of Ventral Visual Processing

In addition to the approaches to build new models discussed above, it is just as important to highlight the importance of methods to more reproducibly and efficiently test and adjudicate among alternative models. These include, for example, benchmarking platforms to collect and maintain all past tests (Schrimpf et al. 2018), methods to pit SMART models against each other to discover controversial stimuli on which their predictions most disagree (Golan et al. 2020), and methods to interpret the results of such tests (Canatar et al. 2023).

4.3. Building SMART Models of Visual Intelligence Beyond Object Recognition

Human visual intelligence is not just object recognition, and it is derived from the entire visual system, not just the ventral stream. Limiting models to just object recognition hinders our understanding of visual intelligence and also will likely not lead to an understanding of the full suite of mechanisms at work in the ventral stream and the rest of the visual system.

The ventral visual stream, often termed the what pathway, has been traditionally associated with object and form representation, while its counterpart, the dorsal stream, often called the where pathway, has been associated with representing spatial location, motion, and guiding actions like grasping. Many recent studies have shown that the functional specializations of these pathways are more complex and often overlap (de Haan & Cowey 2011). In addition, recent developments in computer vision also facilitate incorporating other behavioral tasks, like object detection (Zhao et al. 2019) and monocular depth prediction (Zhao et al. 2020), into models. Beyond recognizing individual objects, our brain processes entire scenes, recognizing actions and interactions of various agents (McMahon & Isik 2023), understanding contexts (Zhang et al. 2020), and making predictions based on the environment and the physics of the world (Bear et al. 2021). Therefore,

SMART models should be developed to understand how these two pathways interact and integrate visual information.

More broadly, to truly understand and model human visual intelligence, we must venture beyond just object recognition and delve into the myriad of other tasks that our visual system performs. This is likely not only to involve what are now mainstream ANN optimization methods, but also to be accelerated by modeling methods that begin with symbolic structure, can generate alternative internal predictions at some level of representation (Lake et al. 2015), and can explicitly manage probabilistic inference in a manner that can scale (Gothoskar et al. 2021). Indeed, the field is now seeing a fusion of such approaches with ANN optimization methods and, when neurally mapped, this will produce new SMART models that will need to be experimentally adjudicated.

4.4. Building and Evaluating Alternative, Cellular-Level Implementations of SMART Models

Many aspects of the known primate brain circuit architecture are not explicitly mapped onto the current SMART models. While it is possible that functional approximations of such motifs are already present in these models in some form, an explicit mapping is definitely missing, rendering these models less interpretable (Kar et al. 2022). These motifs include, but are not limited to, cortico-cortical recurrence, thalamocortical loops, basal ganglia loops, cortical laminar structure and local circuitry, cell types, biophysically grounded dendritic and neuronal models, synaptic dynamics and adaptation, and spiking mechanisms. It is currently unclear how much these structures will turn out to be critical for closing the accuracy gaps in predicting the behaviors supported by the ventral stream or predicting functional neural measurements along the ventral stream (Section 2.1). However, one fundamental principle in neuroscience is that form, encompassing morphology and anatomy, invariably constrains function. By this logic, enhancing SMART models to more closely mirror these recognized anatomical facts will likely bolster their empirical functional accuracy.

The challenge of computationally integrating all of these elements remains formidable. As a result, researchers are taking a piecemeal approach, examining the impact of each omitted or inaccurately represented element individually. Such examples include work on recurrence (Kar et al. 2019, Kubilius et al. 2019, Nayebi et al. 2021, Tang et al. 2018, Zamir et al. 2017) and incorporating cell types (Blauch et al. 2022, Cornford et al. 2020). We note that the endeavor of integrating new neuroscientific components into SMART models can add value in key ways beyond overall improvements in brain alignment against existing measurements. In particular, even if no quantitative empirical accuracy gains are realized, incorporating these components provides routes that could allow for novel perturbation and control tests (see Section 2.2) that might reveal new clinical translation opportunities (further discussed below).

5. POSSIBLE APPLICATIONS OF SMART MODELS

This review highlights the advancement toward brain-mapped ANNs as leading SMART models in object recognition (Section 2), acknowledging their current empirical inaccuracies (Section 3) and limited implementations at fine-grained (e.g., subcellular) levels (type IV; Section 1). As SMART models evolve, we discuss a philosophical issue: their critique as uninterpretable black boxes in neuroscience and artificial intelligence (AI). While these models are fully observable, making the criticism somewhat inappropriate, their complex nature makes intuitive understanding of their behavioral decision processes difficult. In neuroscience, this criticism has resonance because, if the goal is to use SMART models as a proxy for our understanding of the brain's

visual processing, then weaknesses in interpretability seem like limitations. More succinctly, if we succeed in building a digital twin (i.e., a SMART model) but do not fully understand it in all the ways outlined above, how can we say that we understand the brain system that it purports to explain? We expect that theoretical approaches that examine fully observable SMART models, rather than the brain itself, will help our field close some of these gaps (Cohen et al. 2020, Poggio et al. 2020). However, what if this does not—or cannot—happen to our satisfaction?

In this review, we first engage this criticism by logically articulating the concept of mechanistic understanding (Section 1), and we note that many celebrated mechanistic models in neuroscience are subject to similar criticisms (e.g., the Hodgin-Huxley model of action potentials, which is nonlinear and not always intuitively predictable). However, in this section, we put forward another, even more important answer to this criticism: Beyond our field's quest for scientific understanding is a pragmatic goal—to improve the quality of human life. As we outline next (paraphrased from Schrimpf et al. 2020), even difficult-to-interpret SMART models will almost surely be capable of guiding us to new ways to do this.

5.1. Application in Basic Neuroscience Research

In the ventral visual stream, SMART models are already being used to focus experimental resources on the most interesting aspects of brain function that are not yet accurately described. For example, by drawing on the predictive accuracy of these models, neuroscientists can now use them to control individual neurons and entire populations of neurons deep in the visual system via model-synthesized patterns of light applied to retinae (Bashivan et al. 2019, Ponce et al. 2019). Such model-driven stimulus synthesis methods can be used to better adjudicate among alternative SMART models (Golan et al. 2020). Similarly, variants of SMART models that predict image memorability can be used to discover image manipulations that causally affect human memory performance (Goetschalckx et al. 2019).

In another study, SMART models were used to discover that the macaque IT cortical responses are surprisingly sensitively to small, model-guided image perturbations (Guo et al. 2022) and that human category judgments are also surprisingly sensitively to these perturbations (Gaziv et al. 2023). Given the vastness of image space, this previously unknown neurobiology and these previously unknown perceptual sensitivities would have been impossible to discover without these models. Indeed, these discoveries trace back to theoretical and empirical analyses of the SMART models themselves (Goodfellow et al. 2014) and efforts to build new candidate SMART models (Madry et al. 2017).

Stepping back, one can see that the overall research trend here is this: Step 1 is for neuroscientists, cognitive scientists, and computer scientists to work together to transfer the structure of a brain subsystem (which is only partially measurable) into one or more SMART models, where reasonable choices of task and optimization methods help fill in much of the nonmeasurable model structure. In step 2, they then use those now fully observable digital twin SMART models to make and test predictions about that brain subsystem, leading to new scientific discoveries and exposing new model-versus-brain mismatches. Step 3 is to use those models (not the brain itself!) to build a deeper theoretical understanding, which then leads to new SMART models (i.e., a new run of step 1). Repeat the cycle.

5.2. Application in Other Domains

As our field discovers SMART models that are ever more closely aligned to primate brains and primate behavior, gains will naturally follow in several domains.

In AI and computer vision, we will, in fact, be discovering machine systems that, for example, successfully generalize more like humans, are less susceptible to adversarial attacks, and are potentially more energetically efficient. For more general reviews on this topic, we refer the reader to Hassabis et al. (2017).

In direct brain–machine interfaces, sufficiently accurate SMART models of visual processing can be used to determine complex, nonintuitive direct brain stimulation patterns (Azadi et al. 2023, Chen et al. 2020) that could be applied in mid- and high-level visual areas to replicate visual percepts (e.g., in blind individuals).

In mental health, for most brain disorders, the treatment goal is to precisely modulate brain activity in a beneficial way. Going beyond pharmaceuticals (difficult to target precisely) and inserted probes (dangerous and still not precise), accurate SMART models might reveal entirely new treatment possibilities. For instance, these models could direct the synthesis of patterns of light delivered to the retina that predictably and precisely modulate entire populations of neurons deep in the brain at single-neuron resolution to, in turn, beneficially improve cognitive states such as anxiety or depression. To pursue this and other such interventions, a concerted effort is needed to develop and fine-tune models that cater to individuals with unique neurological and behavioral challenges. These refined models, when validated, could revolutionize clinical interventions. Model generators that can explain human variation will be needed to unlock this utility (for example, see Kar 2022). In addition, the more control knobs that a model can engage with, the more potential clinical interventions it can provide, underscoring the imperative to build SMART models that connect to the cellular and molecular components (type IV SMART models) to unlock powerful molecular and genetic interventional toolboxes.

5.3. The Future

We hope the reader will see how these same application themes can—and, we think, will—readily generalize to SMART models of other aspects of cognition, such as audition (Kell et al. 2018), language (Schrimpf et al. 2021), motor planning (Rajalingham et al. 2022), and beyond. The key overall point is that all of the above applications—and myriad others not yet imagined—will get ever better with ever more accurate SMART models, even if we do not fully understand those models in all the ways that we aspire to.

SUMMARY POINTS

1. The past decade has seen ventral stream–inspired deep artificial neural networks (ANNs) emerge that have achieved unprecedented, human-like accuracies on core object recognition tasks.
2. Specific ANNs, once mapped to the brain [sensory-computable, mechanistic, anatomically referenced, and testable (SMART) models], have internal neural representations that surprisingly mimic activity along the primate ventral visual pathway and core object recognition behavioral patterns.
3. Current leading SMART models can be used to synthesize goal-directed stimuli that successfully modulate targeted neural populations in nontrivial ways.
4. Current ANN models do not yet fully capture the neurobehavioral nuances of the ventral visual processing stream and its supported core object recognition behaviors.

FUTURE ISSUES

1. The next step is to create next-generation SMART models that ensure tighter integration with neural and behavioral experiments, leading to corresponding application gains.
2. Researchers should explore and evaluate alternative SMART models that expand beyond visual object recognition.
3. Future research should delve into cellular-level implementations of SMART models, integrating known anatomical details such as cortico-cortical recurrence, cell types, and synaptic dynamics.
4. As the field begins to leverage this new SMART model—based understanding for potential clinical interventions and behavioral modulation, it is crucial to ensure that ethical considerations are at the forefront.

DISCLOSURE STATEMENT

J.J.D. is the current Director of the MIT Quest for Intelligence and a member of the Scientific Advisory Board of the Wu Tsai Institute at Yale and the ARNI NSF Center at Columbia. J.J.D. and K.K. are listed co-inventors on a US patent on methods to use sensory-computable models to modulate brain activity.

ACKNOWLEDGMENTS

K.K. was supported by the Canada Research Chair Program, Simons Foundation Autism Research Initiative (SFARI) grant 967073; a Google Research Award; and the Canada First Research Excellence Funds (VISTA Program). J.J.D. was partially funded by the Office of Naval Research (grants N00014-20-1-2589 and MURI N00014-21-1-2801), the National Science Foundation (grants 2124136), the Simons Foundation (grant 542965), the Semiconductor Research Corporation (SRC), and DARPA. We thank Martin Schrimpf, Carlos Ponce, and Will Xiao for sharing editable drafts of figures and helpful discussions.

LITERATURE CITED

Arend L, Han Y, Schrimpf M, Bashivan P, Kar K, et al. 2018. *Single units in a deep neural network functionally correspond with neurons in the brain: preliminary results*. Tech. Rep., Cent. Brains Minds Mach., Cambridge, MA

Azadi R, Bohn S, Lopez E, Lafer-Sousa R, Wang K, et al. 2023. Image-dependence of the detectability of optogenetic stimulation in macaque inferotemporal cortex. *Curr. Biol.* 33(3):581–88

Barbu A, Mayo D, Alverio J, Luo W, Wang C, et al. 2019. ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 9448–58. Red Hook, NY: Curran Assoc.

Bashivan P, Kar K, DiCarlo JJ. 2019. Neural population control via deep image synthesis. *Science* 364(6439):eaav9436

Bear DM, Wang E, Mrowca D, Binder FJ, Tung HYF, et al. 2021. Physion: evaluating physical prediction from vision in humans and machines. arXiv:2106.08261 [cs.AI]

Blauch NM, Behrmann M, Plaut DC. 2022. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *PNAS* 119(3):e2112566119

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.* 46:e385

Bracci S, Op de Beeck HP. 2023. Understanding human object vision: A picture is worth a thousand representations. *Annu. Rev. Psychol.* 74:113–35

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, et al. 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Comput. Biol.* 15(4):e1006897

Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, et al. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Comput. Biol.* 10(12):e1003963

Canatar A, Feather J, Wakhloo A, Chung S. 2023. A spectral theory of neural prediction and alignment. arXiv:2309.12821 [q-bio.NC]

Chang L, Egger B, Vetter T, Tsao DY. 2021. Explaining face representation in the primate brain using different computational models. *Curr. Biol.* 31(13):2785–95

Chen X, Wang F, Fernandez E, Roelfsema PR. 2020. Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex. *Science* 370(6521):1191–96

Churchland PS, Sejnowski TJ. 1988. Perspectives on cognitive neuroscience. *Science* 242(4879):741–45

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6(1):27755

Cohen U, Chung S, Lee DD, Sompolinsky H. 2020. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* 11(1):746

Cornford J, Kalajdzievski D, Leite M, Lamarquette A, Kullmann DM, Richards BA. 2020. Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. bioRxiv 2020.11.02.364968. **https://doi.org/10.1101/2020.11.02.364968**

Dapello J, Kar K, Schrimpf M, Geary RB, Ferguson M, et al. 2022. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. bioRxiv 2022.07.01.498495. **https://doi.org/10.1101/2022.07.01.498495**

Dapello J, Marques T, Schrimpf M, Geiger F, Cox D, DiCarlo JJ. 2020. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, pp. 13073–87. Red Hook, NY: Curran Assoc.

de Haan EH, Cowey A. 2011. On the usefulness of 'what' and 'where' pathways in vision. *Trends Cogn. Sci.* 15(10):460–66

DiCarlo JJ, Maunsell JH. 2000. Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. *Nat. Neurosci.* 3(8):814–21

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

Dobs K, Martinez J, Kell AJ, Kanwisher N. 2022. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* 8(11):eabl8913

Doshi FR, Konkle T. 2023. Cortical topographic motifs emerge in a self-organized map of object space. *Sci. Adv.* 9(25):eade8187

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929 [cs.CV]

Elsayed G, Shankar S, Cheung B, Papernot N, Kurakin A, et al. 2018. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 3914–24. Red Hook, NY: Curran Assoc.

Fan J, Zeng Y. 2023. Challenging deep learning models with image distortion based on the abutting grating illusion. *Patterns* 4(3):100695

Fel T, Rodriguez Rodriguez IF, Linsley D, Serre T. 2022. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, ed. S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, A Oh, pp. 9432–46. Red Hook, NY: Curran Assoc.

Felleman DJ, Van Essen DC. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1(1):1–47

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. 2013. A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* 16(7):974–81

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

Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC. 1996. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.* 76(4):2718–39

Gattass R, Gross CG, Sandell JH. 1981. Visual topography of V2 in the macaque. *J. Comp. Neurol.* 201(4):519–39

Gattass R, Sousa A, Gross C. 1988. Visuotopic organization and extent of V3 and V4 of the macaque. *J. Neurosci.* 8(6):1831–45

Gaziv G, Lee MJ, DiCarlo JJ. 2023. Robustified ANNs reveal wormholes between human category percepts. arXiv:2308.06887 [cs.CV]

Geiger F, Schrimpf M, Marques T, DiCarlo JJ. 2020. Wiring up vision: minimizing supervised synaptic updates needed to produce a primate ventral stream. bioRxiv 2020.06.08.140111. **https://doi.org/10.1101/2020.06.08.140111**

Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. 2018a. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 [cs.CV]

Geirhos R, Temme CR, Rauber J, Schütt HH, Bethge M, Wichmann FA. 2018b. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 7549–61. Red Hook, NY: Curran Assoc.

Goetschalckx L, Andonian A, Oliva A, Isola P. 2019. GANalyze: toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–53. Piscataway, NJ: IEEE

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

Goodfellow IJ, Shlens J, Szegedy C. 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572 [stat.ML]

Gothoskar N, Cusumano-Towner M, Zinberg B, Ghavamizadeh M, Pollok F, et al. 2021. 3DP3: 3D scene perception via probabilistic programming. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J. Wortman Vaughan, pp. 9600–12. Red Hook, NY: Curran Assoc.

Gross CG, Rocha-Miranda CE, Bender DB. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35(1):96–111

Grossman S, Gaziv G, Yeagle EM, Harel M, Mégevand P, et al. 2019. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* 10(1):4934

Güçlü U, van Gerven MA. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35(27):10005–14

Guo C, Lee M, Leclerc G, Dapello J, Rao Y, et al. 2022. Adversarially trained neural representations are already as robust as biological neural representations. *Proc. Mach. Learn. Res.* 162:8072–81

Hassabis D, Kumaran D, Summerfield C, Botvinick M. 2017. Neuroscience-inspired artificial intelligence. *Neuron* 95(2):245–58

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–78. Piscataway, NJ: IEEE

Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160(1):106–52

Hubel DH, Wiesel TN. 1968. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195(1):215–43

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310(5749):863–66

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

Jazayeri M, Afraz A. 2017. Navigating the neural space in search of the neural code. *Neuron* 93(5):1003–14

Jones H, Grieve K, Wang W, Sillito A. 2001. Surround suppression in primate V1. *J. Neurophysiol.* 86(4):2011–28

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17(11):4302–11

Kar K. 2022. A computational probe into the behavioral and neural markers of atypical facial emotion processing in autism. *J. Neurosci.* 42(25):5115–26

Kar K, DiCarlo JJ. 2021. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron* 109(1):164–76

Kar K, Kornblith S, Fedorenko E. 2022. Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nat. Mach. Intell.* 4(12):1065–67

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 22(6):974–83

Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98(3):630–44

Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Comput. Biol.* 10(11):e1003915

Klindt D, Ecker AS, Euler T, Bethge M. 2017. Neural system identification for large populations separating "what" and "where." In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, ed. I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 3509–19. Red Hook, NY: Curran Assoc.

Kornblith S, Norouzi M, Lee H, Hinton G. 2019. Similarity of neural network representations revisited. *Proc. Mach. Learn. Res.* 97:3519–29

Kriegeskorte N, Mur M, Bandettini PA. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, ed. F Pereira, CJ Burges, L Bottou, KQ Weinberger, pp. 1097–105. Red Hook, NY: Curran Assoc.

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, et al. 2019. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, ed. H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox, R Garnett, pp. 12805–16. Red Hook, NY: Curran Assoc.

Kuhn TS. 1962. *The Structure of Scientific Revolutions*. Chicago: Univ. Chicago Press

Lafer-Sousa R, Conway BR. 2013. Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nat. Neurosci.* 16(12):1870–78

Lake BM, Salakhutdinov R, Tenenbaum JB. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–38

LeCun Y, Bengio Y. 1995. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, ed. MA Arbib, pp. 255–88. Cambridge, MA: MIT Press

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

Lee H, Margalit E, Jozwik KM, Cohen MA, Kanwisher N, et al. 2020. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. bioRxiv 2020.07.09.185116. **https://doi.org/10.1101/2020.07.09.185116**

Levitt JB, Kiper DC, Movshon JA. 1994. Receptive fields and functional architecture of macaque V2. *J. Neurophysiol.* 71(6):2517–42

Logothetis NK, Pauls J, Poggio T. 1995. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5(5):552–63

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

Majaj NJ, Hong H, Solomon EA, DiCarlo JJ. 2015. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* 35(39):13402–18

Margalit E, Lee H, Finzi D, DiCarlo JJ, Grill-Spector K, Yamins DL. 2023. A unifying principle for the functional organization of visual cortex. bioRxiv 2023.05.18.541361. **https://doi.org/10.1101/2023.05.18.541361**

Maunsell JH. 2015. Neuronal mechanisms of visual attention. *Annu. Rev. Vis. Sci.* 1:373–91

Maunsell JH, Treue S. 2006. Feature-based attention in visual cortex. *Trends Neurosci.* 29(6):317–22

McMahon E, Isik L. 2023. Seeing social interactions. *Trends Cogn. Sci.* 27(12):1165–79

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

Miller EK, Gochin PM, Gross CG. 1991. Habituation-like decrease in the responses of neurons in inferior temporal cortex of the macaque. *Vis. Neurosci.* 7(4):357–62

Nayebi A, Sagastuy-Brena J, Bear DM, Kar K, Kubilius J, et al. 2021. Goal-driven recurrent neural network models of the ventral visual stream. bioRxiv 2021.02.17.431717. **https://doi.org/10.1101/2021.02.17.431717**

Ngo J, Sankaranarayanan S, Isola P. 2023. *Is CLIP fooled by optical illusions?* Tiny Pap., Int. Conf. Learn. Rep., N.p. **https://openreview.net/forum?id=YdGkE4Ugg2C**

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. 2014. A toolbox for representational similarity analysis. *PLOS Comput. Biol.* 10(4):e1003553

Nuthmann A. 2017. Fixation durations in scene viewing: modeling the effects of local image features, oculomotor parameters, and task. *Psychon. Bull. Rev.* 24(2):370–92

Op De Beeck H, Vogels R. 2000. Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426(4):505–18

Op de Beeck H, Wagemans J, Vogels R. 2001. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4(12):1244–52

Parvizi J, Jacques C, Foster BL, Withoft N, Rangarajan V, et al. 2012. Electrical stimulation of human fusiform face-selective regions distorts face perception. *J. Neurosci.* 32(43):14915–20

Pasupathy A, Connor CE. 1999. Responses to contour features in macaque area V4. *J. Neurophysiol.* 82(5):2490–502

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. 2011. A molecular phylogeny of living primates. *PLOS Genet.* 7(3):e1001342

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

Phillips RR, Malamut BL, Bachevalier J, Mishkin M. 1988. Dissociation of the effects of inferior temporal and limbic lesions on object discrimination learning with 24-h intertrial intervals. *Behav. Brain Res.* 27(2):99–107

Pinto N, Doukhan D, DiCarlo JJ, Cox DD. 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLOS Comput. Biol.* 5(11):e1000579

Poggio T, Banburski A, Liao Q. 2020. Theoretical issues in deep networks. *PNAS* 117(48):30039–45

Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* 177(4):999–1009

Popivanov ID, Jastorff J, Vanduffel W, Vogels R. 2014. Heterogeneous single-unit selectivity in an fMRI-defined body-selective patch. *J. Neurosci.* 34(1):95–111

Popper KR. 1934. *The Logic of Scientific Discovery*. Berlin: Julius Springer

Pospisil DA, Pasupathy A, Bair W. 2018. "Artiphysiology" reveals V4-like shape tuning in a deep network trained for image classification. *eLife* 7:e38242

Potter MC. 1976. Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. Learn. Mem.* 2(5):509–22

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

Rajalingham R, Piccato A, Jazayeri M. 2022. Recurrent neural networks with explicit representation of dynamic latent variables can mimic behavioral patterns in a physical inference task. *Nat. Commun.* 13(1):5865

Rajalingham R, Schmidt K, DiCarlo JJ. 2015. Comparison of object recognition behavior in human and monkey. *J. Neurosci.* 35(35):12127–36

Rajalingham R, Sorenson M, Azadi R, Bohn S, DiCarlo JJ, Afraz A. 2021. Chronically implantable led arrays for behavioral optogenetics in primates. *Nat. Methods* 18(9):1112–16

Ratan Murty NA, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N. 2021. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* 12(1):5540

Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, et al. 2019. A deep learning framework for neuroscience. *Nat. Neurosci.* 22(11):1761–70

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

Ringach DL, Shapley RM, Hawken MJ. 2002. Orientation selectivity in macaque V1: diversity and laminar dependence. *J. Neurosci.* 22(13):5639–51

Rossion B, Taubert J. 2019. What can we learn about human individual face recognition from experimental studies in monkeys? *Vis. Res.* 157:142–58

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature* 323(6088):533–36

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115:211–52

Rust NC, Schwartz O, Movshon JA, Simoncelli EP. 2005. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* 46(6):945–56

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, et al. 2021. The neural architecture of language: integrative modeling converges on predictive processing. *PNAS* 118(45):e2105646118

Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv 407007. **https://doi.org/10.1101/407007**

Schrimpf M, Kubilius J, Lee MJ, Murty NAR, Ajemian R, DiCarlo JJ. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108(3):413–23

Serre T, Riesenhuber M. 2004. *Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex*. Rep., Comput. Sci. Artif. Intell. Lab., Mass. Inst. Technol., Cambridge

Sexton NJ, Love BC. 2022. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Sci. Adv.* 8(28):eabm2219

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–26. Piscataway, NJ: IEEE

Tanaka K. 1996. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19:109–39

Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, et al. 2018. Recurrent computations for visual pattern completion. *PNAS* 115(35):8835–40

Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. *Nature* 381(6582):520–22

Tsao DY, Freiwald WA, Tootell RB, Livingstone MS. 2006. A cortical region consisting entirely of face-selective cells. *Science* 311(5761):670–74

Ungerleider LG, Mishkin M, et al. 1982. Two cortical visual systems. In *Analysis of Visual Behavior*, ed. DJ Ingle, MA Goodale, RJW Mansfield, pp. 549–86. Cambridge, MA: MIT Press

Vogels R. 2022. More than the face: representations of bodies in the inferior temporal cortex. *Annu. Rev. Vis. Sci.* 8:383–405

Walker EY, Sinz FH, Cobos E, Muhammad T, Froudarakis E, et al. 2019. Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.* 22(12):2060–65

Wolff SB, Ölveczky BP. 2018. The promise and perils of causal circuit manipulations. *Curr. Opin. Neurobiol.* 49:84–94

Xiao W, Kreiman G. 2020. XDream: finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Comput. Biol.* 16(6):e1007973

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 C, DiCarlo JJ. 2013. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in Neural Information Processing Systems 26 (NeurIPS 2013)*, ed. CJ Burges, L Bottou, M Welling, Z Ghahramani, KQ Weinberger, pp. 3093–101. Red Hook, NY: Curran Assoc.

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

Zamir AR, Wu TL, Sun L, Shen WB, Shi BE, et al. 2017. Feedback networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1308–17. Piscataway, NJ: IEEE

Zeiler MD, Fergus R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014*, pp. 818–33. Berlin: Springer

Zhang M, Tseng C, Kreiman G. 2020. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12985–94. Piscataway, NJ: IEEE

Zhang Y, Meyers EM, Bichot NP, Serre T, Poggio TA, Desimone R. 2011. Object decoding with attention in inferior temporal cortex. *PNAS* 108(21):8850–55

Zhao C, Sun Q, Zhang C, Tang Y, Qian F. 2020. Monocular depth estimation based on deep learning: an overview. *Sci. China Technol. Sci.* 63(9):1612–27

Zhao ZQ, Zheng P, St Xu, Wu X. 2019. Object detection with deep learning: a review. *IEEE Trans. Neural Netw. Learn. Syst.* 30(11):3212–32

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, et al. 2021. Unsupervised neural network models of the ventral visual stream. *PNAS* 118(3):e2014196118