How does the primate ventral visual stream causally support core object recognition?

by

Rishi Rajalingham

B.Eng., McGill University (2010)
M.Eng., McGill University (2012)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Brain and Cognitive Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author ..... Signature redacted .....
Department of Brain and Cognitive Sciences
June 18, 2018

Certified by... Signature redacted .....
James J. DiCarlo
Peter de Florez Professor of Neuroscience,
Head, Department of Brain and Cognitive Sciences
Thesis Supervisor

Accepted by .. Signature redacted .....
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience,
Director of Graduate Education for Brain and Cognitive Sciences

How does the primate ventral visual stream causally support core object recognition?

by

Rishi Rajalingham

Submitted to the Department of Brain and Cognitive Sciences
on June 18, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Brain and Cognitive Sciences

Abstract

Primates are able to rapidly, accurately and effortlessly perform the computationally difficult visual task of invariant object recognition — the ability to discriminate between different objects in the face of high variation in object viewing parameters and background conditions. This ability is thought to rely on the ventral visual stream, a hierarchy of visual cortical areas culminating in inferior temporal (IT) cortex. In particular, decades of research strongly suggests that the population of neurons in IT supports invariant object recognition behavior. However, direct causal evidence for this decoding hypothesis has been equivocal to date, especially beyond the specific case of face-selective sub-regions of IT. This research aims to directly test the general causal role of IT in invariant object recognition. To do so, we first characterized human and macaque monkey behavior over a large behavioral domain consisting of binary discriminations between images of basic-level objects, establishing behavioral metrics and benchmarks for computational models of this behavior. This work suggests that, in the domain of basic-level core object recognition, humans and monkeys are remarkably similar in their behavioral responses, while leading models of the visual system significantly diverge from primate behavior. We then reversibly inactivated individual, millimeter-scale regions of IT via injection of muscimol while monkeys performed several interleaved binary object discrimination tasks. We found that inactivating different millimeter-scale regions of primate IT resulted in different patterns of object recognition deficits, each predicted by the local region's neuronal selectivity. Our results provide causal evidence that IT directly underlies primate object recognition behavior in a topographically organized manner. Taken together, these results establish quantitative experimental constraints for computational models of the ventral visual stream and object recognition behavior.

Thesis Supervisor: James J. DiCarlo
Title: Peter de Florez Professor of Neuroscience,
Head, Department of Brain and Cognitive Sciences

Acknowledgments

This thesis is a reflection of six years of unique experiences that were made possible by the contributions of many people; I take this opportunity to thank them.

First and foremost, I have been fortunate to learn under the guidance of Jim DiCarlo, a superhuman scientist. I'm immensely grateful to him for his unconditional support and generosity, and for sharing his inspired vision and staunch commitment to scientific rigour. Working with Jim has fueled my desire to question and to learn. I'm especially grateful for the invaluable academic environment he created, and to all members of DiCarlo lab, who have made this a rich experience. In particular, I am indebted to the broad and deep curiosity of Arash Afraz, the diligence and wisdom of Elias Issa, the generosity of Kailyn Schmidt, the virtuosity of Shay Ohayon, and the acumen of Daniel Yamins.

I am thankful to my thesis committee members — Nancy Kanwisher, Mehrdad Jazayeri, and Roozbeh Kiani — for their help throughout the program, and for their diversity and mastery of thought.

To Manto, Zico, Picasso, Yolo and Espresso.

To friends beyond the lab — Diego Vargas, Jasmin Imsirovic, Pedro Tsividis, Alex Kell, Sam Norman-Haignere, Jorie Koster-Hale, Or Shemesh, Greg Stacey, thank you for your laughter and your reflection.

To my family, for imparting the value of education; to my mother, for her strength.

Finally, to Sonya, for reminding me of the value of chasing dreams.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

1.1 Visual object recognition

A conjecture: the computational goal of the brain is to perceive the world and act upon it. In the process, the brain gives rise to the mind, an emergent property that is both intractably ill-defined and infuriatingly interesting. To gain understanding of the origins and mechanisms of the mind—understanding that goes beyond the limits of introspection—it has been helpful to first reduce it into operationally defined observable phenomena. To this end, our conscious subjective experience can be reduced into observable behaviors encompassed in domains of perception, memory, decision-making, etc., which together achieve the aforementioned goal of perceiving the world and acting upon it. Visual object recognition is one of those behaviors.

Visual object recognition refers to the ability to assign labels (e.g. identity, category, etc.) to objects in visual stimuli. While this behavior may seem arbitrary and contrived, the motivation to study the neural mechanisms underlying it are two-fold. In primates (including humans), this ability—and more broadly, the ability to perceive affordances in environments [Gibson, 1979]—is crucial for survival and well-being. Threat detection, resource detection, navigation, and social interactions are all behaviors that depend significantly on this perceptual ability. As such, primates have evolved to rapidly and accurately recognize objects in visual images. Interestingly, this behavior introspectively appears to be reflexive or relatively effortless,

which can mislead one to infer that the underlying neural computations are relatively straight-forward. An infamous early attempt at solving this problem is the *Summer Vision Project* [Papert, 1966], which was "an attempt to use our summer workers effectively in the construction of a significant part of a visual system. ... The final goal is OBJECT IDENTIFICATION which will actually name objects by matching them with a vocabulary of known objects." Needless to say, this problem was not solved that one summer in 1966, and five decades of significant efforts on this problem have only recently yielded progress in approaching human abilities in some visual object recognition tasks [Krizhevsky et al., 2012]. Indeed, the problem of object recognition is inherently ill-posed, given that any three-dimensional object can project nearly infinitely many 2D images on the retina, under variations in pose, position, size, articulation, illumination, and background context. Figure 1-1A illustrates this problem for an example task (discriminating between a car and non-car objects, in the face of viewpoint variability). *Invariance* to such identity-preserving image transformations is thought to be the computational crux of this problem, and this invariant object recognition has traditionally been challenging for artificial vision systems [Ullman, 2000, Pinto et al., 2008]. Taken together, the ecological importance of this visual ability and the relative ease with which it is solved by the primate brain motivate a quantitative understanding of the neural mechanisms underlying primate invariant object recognition behavior.

For the purpose of this work, we define quantitative understanding of the neural mechanisms underlying visual object recognition as uncovering a model that recapitulates observed all relevant neural and behavioral phenomena in this behavioral domain. The choices for what is "relevant" are numerous, open to debate, and likely depend on the end goals of this understanding (e.g. building artificial intelligence, restoring or augmenting human abilities, etc.). To gain traction towards this ambitious goal, we focus on a subset of invariant recognition behaviors that capture the computational crux described above. To this end, *core invariant object recognition* is operationally defined as the ability to identify objects under high view uncertainty in visual images in the central visual field during a single, natural viewing fixation

[DiCarlo et al., 2012]. We further restricted our behavioral domain to discriminations between basic-level object categories [Rosch et al., 1976]. This reduced behavioral domain aims to captures the primates' perceptual abilities at a glance, a coarse yet surprisingly rich perceptual experience that is a foundation for more sophisticated visual abilities. We do not claim this to be an exhaustive characterization of all possible visual object recognition behaviors, but rather a good starting point for that greater goal. With this operationally defined behavioral domain in hand, we ask: how does the primate brain support basic-level core object recognition behavior?

1.2 The ventral visual stream

Early lesion studies in humans and monkeys implicate the *ventral visual stream* in object recognition behavior [Holmes and Gross, 1984, Horel et al., 1987, Biederman et al., 1997]. Given remarkable anatomical and functional homologies between monkeys and humans [Mantini et al., 2012, Orban et al., 2004], much of the literature reviewed here examines the ventral visual stream in macaque monkeys. The ventral visual stream consists of a series of visual cortical areas that are located along the ventral surface of the macaque monkey brain [Miyashita, 1993, Gross, 1994, Rolls, 2000]. Analogously, the *dorsal* visual stream consists of corresponding cortical series along the dorsal surface, and these two streams are thought to support distinct visual behaviors, caricatured as perception for recognition (ventral stream) and for guidance of actions (dorsal stream) [Goodale and Milner, 1992].

The ventral visual stream, schematized in Figure 1-1B, consists of primary visual cortex (V1), and subsequent extrastriate ventral visual cortical areas V2, V4 and inferior temporal (IT) cortex. When light impinges on the retina, this visual information is transduced into neuronal activity and relayed via the lateral geniculate nucleus (LGN) of the thalamus to these visual cortical areas. While this cortical network is highly recurrent, it can be approximated as a stacked hierarchy of cortical regions (V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ IT) [Felleman and Van Essen, 1991]. The visual input is transformed at each of these cortical stages, creating subsequently more complex visual

representations. Individual neurons in a given cortical area are thought to compute local image features, and together tile the visual field. Across cortical regions, neurons reflect increasingly more explicit representations of behaviorally relevant image properties [Rust and DiCarlo, 2010]. For instance, neurons in V1 appear to compute relatively simple edge-like local features parameterized by local contrast, spatial frequency and orientation [Hubel and Wiesel, 1962, Hubel and Wiesel, 1968]; V2 and V4 are thought to compute conjunctions of these, exhibiting selectivity to complex shapes [Hegdé and Van Essen, 2000] and curvatures [Pasupathy and Connor, 2002]; neurons in IT cortex, which exhibit remarkable tolerance to changes in viewing parameters (e.g. position, scale, and pose), appear to be selective to particular object categories [Bruce et al., 1981, Logothetis et al., 1995].

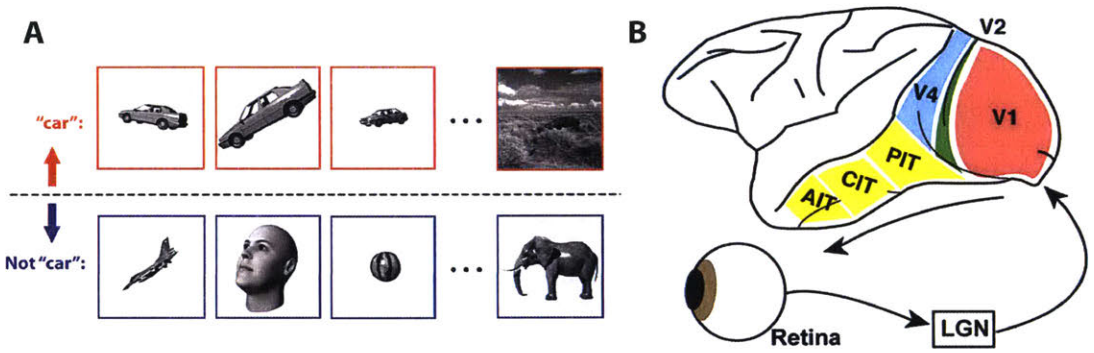

Figure 1-1: (a) Schematic illustrating the computational crux of object recognition. For an example task (discriminating between a car and non-car objects), identity-preserving variability in viewpoint parameters such as position, pose and size leads to drastically different visual inputs. (b) The ventral visual stream, schematized as a hierarchy of visual cortical areas (V1, V2, V4, IT). (Both panels are adapted from [DiCarlo et al., 2012]).

We refer to the question of how neuronal responses are produced from the external visual input as the *encoding* problem. A key goal in systems neuroscience is to construct encoding models, computational models that predict the activity of neurons in response to any/all visual input. To date, the phenomena described above have constrained encoding models of the ventral stream to a class of hierarchical deep convolutional neural networks (DCNNs) that reflect local features tiling the visual

field (via a convolutional structure) and that are hierarchically computed with operations that mimic tolerance and selectivity building (conjunctions and pooling). DCNNs are currently the leading encoding models of the ventral visual stream [Seibert et al., 2016, Cadieu et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014, Guclu and van Gerven, 2015, Cichy et al., 2016, Hong et al., 2016, Seibert et al., 2016, Cadena et al., 2017, Wen et al., 2017].

We refer to the question of how neuronal responses are read out to support behavior as the *decoding* problem. As stated above, decades of research suggest that the ventral visual stream, and in particular IT cortex, is necessary for object recognition behavior [Holmes and Gross, 1984, Horel et al., 1987, Biederman et al., 1997, Logothetis and Sheinberg, 1996, Tanaka, 1996, Rolls, 2000, DiCarlo et al., 2012]. This decoding hypothesis is qualitatively supported by observations of individual neurons' selectivity to object identity and category, with remarkable tolerance to nuisance viewpoint parameters [Kobatake and Tanaka, 1994, Ito et al., 1995, Logothetis et al., 1995, Booth and Rolls, 1998, DiCarlo et al., 2012]. More quantitatively, this decoding hypothesis is supported by observations that a linear readout from the population of neurons in IT not only matches overall primate behavioral performance [Hung et al., 2005, Zhang et al., 2011] but also predicts primate behavioral patterns of performance across different object recognition tasks [Sheinberg and Logothetis, 1997, de Beeck et al., 2001, Majaj et al., 2015], suggesting that IT is a good neural correlate of primate recognition behavior.

1.3 Statement of problem

Decades of research have yielded key insights into how visual stimuli are represented in the brain (encoding problem) and how these representations may support object recognition behavior (decoding problem). Here, our goal was to uncover if and how activity in IT *causally* supports basic-level core object recognition behavior.

1.3.1 Inferring causal dependencies

We first review the definition of inferred causal and correlational dependencies, illustrated in Figure 1-2 [Jazayeri and Afraz, 2017]. One can experimentally measure associations between phenomena X, Y. Associations measured without any intervention (specifically, without experimental randomization of one variable, X) are referred to as correlational dependencies (see Figure 1-2A). Such associations are consistent with a causal link between the measured phenomena (i.e. X causes Y), but could also reflect epiphenomenal processes in the absence of causation, e.g. whereby a third confounding variable drives both measured (i.e. Z causes both X and Y). Recent research in several behavioral domains has exposed discrepancies between correlational dependencies and directly tested causal dependencies [Katz et al., 2016, Liu and Pack, 2017], suggesting potential epiphenomenal mechanisms and highlighting the need to directly test causal hypotheses.

In contrast, causal dependencies can be inferred by linking a dependent variable to an experimentally controlled variable (see Figure 1-2B,C). Note that the distinction between correlation and causal is not simply whether there is experimental control of a variable, but rather what dependence is inferred. For example, direct manipulation of visual input to the retinae while measuring both neuronal activity in IT and behavior allows one to infer causal dependencies between visual inputs and neuronal activity in IT, as well as between visual inputs and behavior, but not between neuronal activity in IT and behavior.

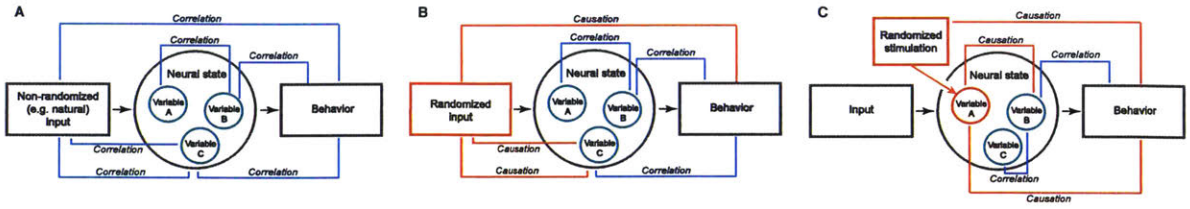

Figure 1-2: Adapted from [Jazayeri and Afraz, 2017]: schematic of three different experiments inferring correlational and causal dependencies between measured variables. Blue arrows correspond to correlational dependencies between two variables, where neither variable is experimentally controlled. Red arrows correspond to inferred causal dependencies, where a dependent variable is linked to an experimentally controlled variable. In (c), the experimentally controlled variable (randomized stimulation) is assumed to be equivalent to the internal variable A.

An important caveat to this definition is that inferred causal dependences often link a dependent variable (Y) to a latent variable (X') that is assumed to be equivalent to the experimentally controlled variable (X). For example, in Figure 1-2C, the experimentally manipulated variable (randomized stimulation) is assumed to be equivalent to the latent variable (A), thus enabling the inference of a causal dependence between A and behavior. However, if the randomized stimulation was not specific to randomizing variable A, or was not successful at randomizing variable A, such an inference would no longer be supported. This caveat is especially important given the crude nature of our neural perturbation tools (e.g. electrical micro-stimulation of clusters of neurons can cause inadvertent stimulation of axons of passage), and suggests that inferred dependencies fall on a continuum rather than two discrete categories of correlational and causal. Thus, we propose the following definition: the confidence of inferred causal dependencies between measured phenomena X', Y depends on the (estimated) equivalence between the experimentally controlled variable X and the inferred causal variable X'.

1.3.2 The causal role of IT in core object recognition

We now apply this framework to testing the following decoding hypothesis: IT cortex is a necessary part of the brain's neural network that underlies core recognition be-

havior — or, stated in other words, core object recognition behavior causally depends on IT cortex. As described above, it has been shown that the population of neurons in IT is a good neural correlate of primate recognition behavior [Sheinberg and Logothetis, 1997, de Beeck et al., 2001, Majaj et al., 2015]. In these experiments, the experimentally controlled variable (visual images shown to the retinae) is not equivalent to neuronal activity in IT. Indeed, while visual images may reliably drive IT activity, they likely drive other confounding variables (e.g. any other visually driven brain area). Thus, inferred dependencies from such an experiment are consistent with, but do not unequivocally support a causal dependence of the measured behavior on neural activity in IT.

To date, the most successful direct manipulations of IT have exclusively targeted at millimeter-scale clusters of face-selective neurons in IT [Afraz et al., 2006, Afraz et al., 2015, Moeller et al., 2017, Sadagopan et al., 2017]. These results suggest that these IT sub-regions are necessary for at least some basic- and subordinate-level face recognition behaviors. However, results from direct manipulations of IT in general visual recognition behavior have been equivocal at best. Lesions of IT sometimes suggest the necessity of IT for visual behaviors [Cowey and Gross, 1970, Manning, 1972, Holmes and Gross, 1984, Biederman et al., 1997, Buffalo et al., 2000] but the resulting behavioral deficits are often contradictory (with often no lasting visual deficits) [Dean, 1974, Huxlin et al., 2000] and at best modest (e.g. 10-15% drop in performance for large-scale bilateral removal of IT when a complete loss of performance would have been 40%) [Horel et al., 1987, Matsumoto et al., 2016]. Thus, it is still unclear if IT is necessary for general core object recognition behavior. Moreover, even if IT cortex is indeed necessary for all core object recognition tasks, it is unclear how that assumed causal role is organized. To answer these questions, we tackle a number of specific aims, summarized below.

1.3.3 Organization of thesis

Aim 1: Behavioral benchmark of human core object recognition behavior.

In this aim, we set out to establish systematic behavioral benchmarks on which to test models of human object recognition. In Chapter 2, we established a scalable behavioral paradigm for testing the object recognition abilities of humans and monkeys on hundreds of different object discrimination tasks. Using this behavior, we were able to systematically benchmark the macaque monkey as a model of human visual processing, justifying this animal model for studying high level vision. In Chapter 3, we significantly scaled up our behavioral experiments to test the limits of state-of-the-art DCNNs. We systematically compared the behavioral responses of these models with the behavioral responses of humans and monkeys, at the resolution of individual images, over thousands of experimental conditions. Using high-resolution behavioral metrics, we found that all tested ANN models significantly diverged from primate behavior. Going forward, these high-resolution, large-scale primate behavioral benchmarks could serve as direct guides for discovering better ANN models of the primate visual system.

Aim 2: Test the causal role of IT in core object recognition behavior.

In Chapter 4, we reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of IT via local injection of muscimol while monkeys performed a battery of binary core object discrimination tasks, interleaved trial-by-trial. First, our results provide direct causal evidence for the role of IT in basic-level core object recognition. Going beyond a qualitative answer, we found that inactivating different millimeter-scale regions of primate IT resulted in different patterns of object recognition deficits, each predicted by the local region's neuronal selectivity. To the best of our knowledge, this is the first study to demonstrate the necessity of IT cortex for a wide range of general core object recognition behaviors with behaviorally critical topographic organization.

Aim 3: Novel technology for optogenetic experiments in primates

In Chapter 5, we tested a novel tool for optogenetic experiments in primates, a chronically implantable array of LEDs. Chronic tools such as this may be promising avenues for high-throughput behavioral experiments with time-delimited perturbation of neural activity. We first characterized the LED arrays' photometric properties for use in-vivo. Following this, we tested the LED arrays in two different behavioral experiments, each testing the causal role of a visual cortical area in an established behavioral task. Our data provide a report of preliminary findings with guides for improvements, both technological and experimental.

Chapter 2

Comparison of Object Recognition Behavior in Human and Monkey

To understand the neural mechanisms underlying high-level vision in humans, it has been necessary to study various animal models, where fine grained mechanistic access is available. To this end, the rhesus monkey is the most widely used animal model of human visual processing. However, it is not known if invariant visual object recognition behavior is quantitatively comparable across monkeys and humans. To address this question, we first defined and systematically compared the object recognition behavior of monkeys with that of humans. To date, several mammalian species have shown promise as animal models for studying the neural mechanisms underlying high-level visual processing in humans. In light of this diversity, making tight comparisons between non-human and human primates is particularly critical to further the field's goal of translating knowledge gained from animal models to humans. To the best of our knowledge, this[1] is the first systematic attempt at comparing a high-level visual behavior of humans and macaque monkeys.

[1]The contents of this chapter are adapted from a published journal article [Rajalingham et al., 2015].

2.1 Introduction

Humans are able to rapidly, accurately and effortlessly perform the computationally difficult visual task of invariant object recognition — the ability to discriminate between different objects in the face of high variation in object viewing parameters and background conditions [DiCarlo et al., 2012]. However, it is still unclear how the human brain supports this behavior. To uncover the neuronal mechanisms underlying human visual processing, it has been necessary to study various animal models including non-human primates, felines and rodents [Hubel and Wiesel, 1962, Van Essen, 1979, Zoccolan et al., 2009]. In particular, the rhesus macaque monkey, an Old World primate that diverged from humans approximately 25 million years ago [Kumar and Hedges, 1998], is one of the most widely used animal models of high-level human visual perception [Mishkin et al., 1983, Tanaka, 1996, Minamimoto et al., 2010, Kravitz et al., 2011, Grill-Spector and Weiner, 2014]. There exist strong anatomical and functional correspondences of visual cortical areas between humans and monkeys [Orban et al., 2004, Mantini et al., 2012, Miranda-Dominguez et al., 2014]. Thus, it has long been assumed that high-level visual behaviors and underlying neural substrates are comparable between monkey and human. However, humans have capacities not found in monkeys, and their brains differ in important ways [Passingham, 2009]. To date, the limits of the similarity in high-level visual behaviors of macaques and humans are unknown as no effort has been made to systematically compare rhesus macaque monkeys with humans in invariant object recognition. In light of recent work showing that rodent models of visual processing display the qualitative ability to perform invariant shape discrimination [Zoccolan et al., 2009], making tight, quantitative comparisons between monkeys and humans is especially critical in determining the best use of non-human primates to further the field's goal of translating knowledge gained from animal models to humans.

To do this, we here systematically compared the behavior of two macaque monkeys with that of normal human subjects on an invariant object recognition paradigm.

Our goal was to make direct measurements of object recognition ability, over a very large number of recognition tasks, always under conditions of high object view variation (a.k.a. "invariant" object recognition). We focused on "core invariant object recognition" —rapid and reliable recognition during a single, natural viewing fixation [DiCarlo and Cox, 2007, DiCarlo et al., 2012], operationalized as images presented in the central 10° of the visual field for durations under 200ms. We further restricted our behavioral domain to "basic-level" object categories, as defined by [Rosch et al., 1976]. We do not claim this to be an exhaustive characterization of all possible visual object recognition behaviors, but rather a good starting point for that greater goal. Monkeys easily learn such tasks and, after testing 276 such object recognition tasks, our results show that rhesus monkey and human behavior are largely indistinguishable, and that both species are easily distinguishable from low-level visual representations asked to perform the same 276 tasks. These results show that rhesus monkeys are a very good — and perhaps quantitatively exact — model of human invariant object recognition abilities, and they are consistent with the hypothesis that monkeys and humans share a common neural representation that directly underlies those abilities.

2.2 Results

As motivated in the Introduction, our primary goal was to directly compare the behavioral abilities of humans and monkeys over a large battery of basic-level invariant object discrimination tasks. Discrimination tasks spanned a set of 24 basic-level objects; Figure 2-1 shows a complete list of all 24 objects, with two example images (out of hundreds) per object. Each invariant object discrimination task consisted of a binary match-to-sample, where subjects were forced to categorize naturalistic synthetic images, of objects presented under high view uncertainty and on a randomized background, (see Figure 2-2).

In sum, the comparisons presented here (Figs. 3 and 4) are based on data obtained from a pool of 605 human subjects (69,000 total trials) and two monkey subjects

Figure 2-1: Two example images for each of the 24 basic-level objects, sampled from the test set (each row corresponds to a group of eight objects). To enforce true object recognition behavior (rather than image matching) and tackle the invariance problem, we generated thousands of naturalistic images, each with one foreground object, by rendering a 3D model of each object with randomly-chosen viewing parameters (2D position, 3D rotation and viewing distance) and placing that foreground object view onto a randomly-chosen, natural image background

(106,844 total trials; see Methods). The monkey data pool only includes behavioral trials collected after the monkey learned all 24 objects. This corresponds to a total of 250 human behavioral trials for each of the 276 tasks, and a total of 362-417 monkey trials for each task.

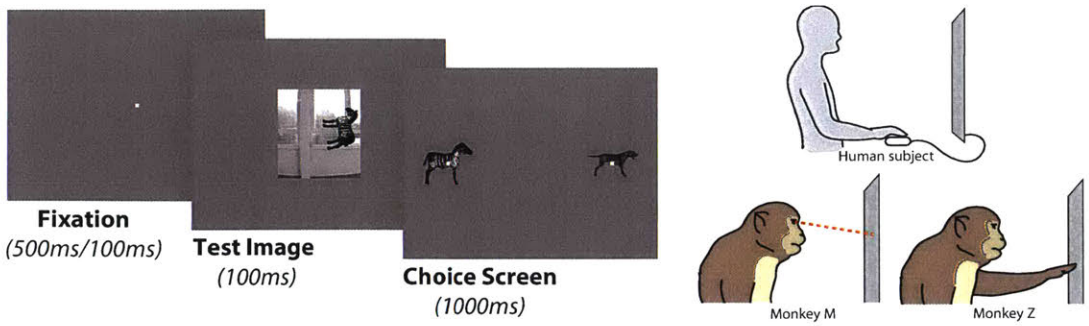

Figure 2-2: Behavioral paradigm (for Monkey M). Each trial was initiated when the monkey held gaze fixation on a central fixation point for 200ms, after which a square test image (spanning 6° of visual angle) appeared at the center of gaze for 100ms. Immediately after extinction of the test image, two choice images, each displaying the canonical view of a single object with no background were shown to the left and right (see Methods). Test and choice images are shown to scale. The monkey was allowed to freely view the response images for up to 1500ms, and responded by holding fixation over the selected image for 700ms. Monkey Z performed the exact same tasks, but used touch to initiate trials and indicate its choice (see Methods). Successful trials were rewarded with juice, and incorrect choices resulted in time-outs of 1.5 to 2.5 seconds.

2.2.1 Learning

As described in the Methods, monkeys rapidly learned each new object. While we do not know the humans' or the monkeys' prior lifetime experiences with these objects or related image statistics, we reasoned that, following monkey training, both species might be in a comparable lifetime training regime. To assess this, we examined the effects of experience in both humans and monkeys by tracking the performance of individual human and monkey subjects as a function of the number of exposures to each object. Figure 2-3A,B show each monkey's performance, relative to the human pool, when presented with 16 novel objects; both monkeys were able to reach high performance relatively quickly (˜$1000 - 2000$ image presentations). Figure 2-3C directly compares both monkeys to individual humans subjects on the exact same tasks ($n = 15$ human subjects with sufficient longitudinal data). Unlike monkeys, individual human subjects initially behaved at a high level of performance, and exhibited no increase in performance as a function of exposures to objects, suggesting that humans have prior experience with similar objects.

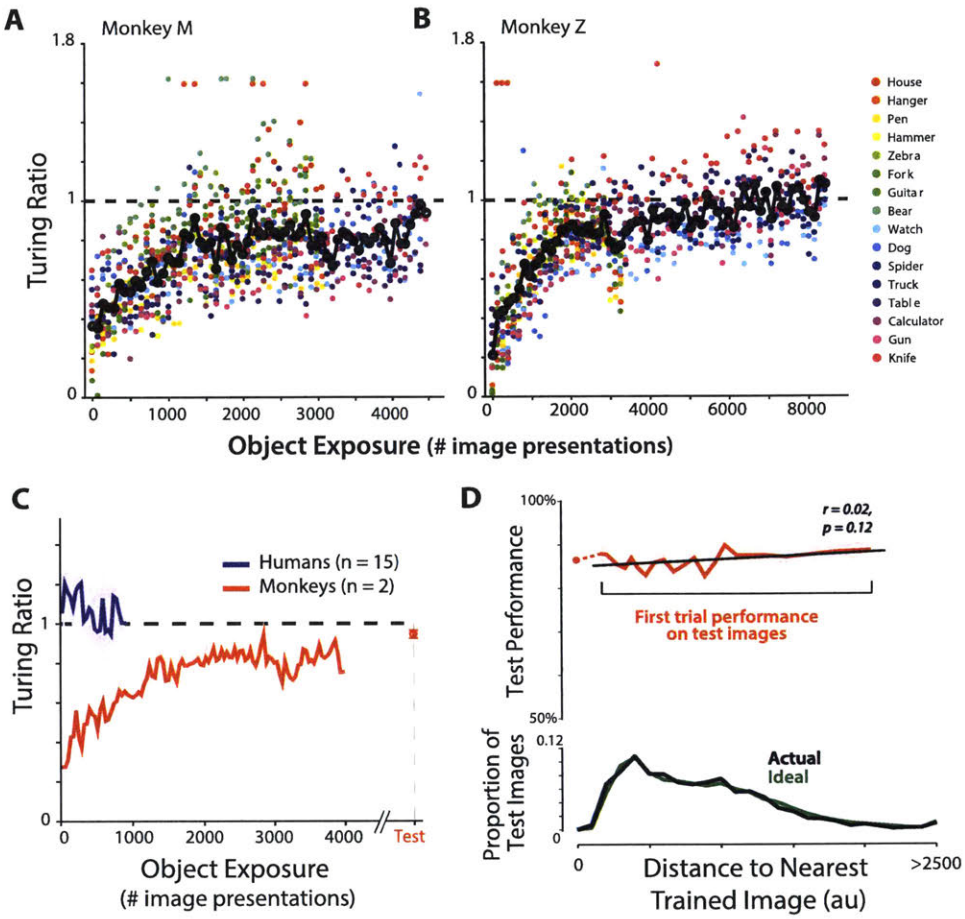

Figure 2-3: Performance relative to the human pool for 16 objects (shown as colored dots; the mean over objects is shown as a solid black line) for monkey M and P. C) Average performance relative to the human pool of two monkeys and 15 unique individual humans subjects with sufficient longitudinal data on the same tasks (mean ± SE over subjects). Monkeys rapidly learned each new object, while humans performed at a high initial performance, and exhibited no change in performance as a function of (unsupervised) experience with the objects. The TEST marker indicates monkeysÕ relative performance on held-out test images, following all behavioral training. D) Top panel: Generalization to novel images. Pooling data from both monkeys, the first-trial performance of 2400 test images (mean ± SE) versus the corresponding Euclidean pixel distance to the nearest training image; black line denotes linear regression. The overall performance, including all subsequent exposures to test images, is shown on the left (at zero distance). Bottom panel: Overlap between training and test image sets. The distribution of distances of test images to the nearest trained image is shown relative to actual training images (black line), and to ÒidealÓ generalization surrogate training images (green line).

Following training, monkeys maintained high performance when tested on novel images of these previously learned objects (see Figure 2-3C, "TEST" marker). Importantly, this generalization was immediate, with comparable high performance on the very first trial of a new image for both monkeys. Furthermore, the generalization performance did not depend on the similarity of test images to previously seen training images (see Figure 2-3D top panel, Methods). Indeed, monkeys maintained high behavioral performance even for the subset of test images that were nearly as far from the training images as they would have been if we had completely restricted training with each single axis of variation (see Methods). Finally, subsequent exposures to these test images did not further increase behavioral performance (see Figure 2-3D, zero distance marker). Taken together, these results suggests that monkeys were not simply memorizing previously learned images, and that they could generalize to significantly different images of the same object, even when the training images only sparsely sample the view space of the object. Importantly, while it is impossible to guarantee or expect that humans and monkeys have identical lifetime experience, we find that, once monkeys were trained, further experience had no observable effect on the behavioral performance of either species.

2.2.2 Human consistency

The difficulty of each of the 276 invariant object discrimination tasks was determined by measuring the unbiased performance (d') of monkeys/humans. That performance metric is shown for all 276 tasks in Figure 2-4A. We refer to these data as the "pattern of behavioral performance" for pooled human (605 subjects, 69,000 trials) and pooled monkey (2 subjects, 106,844 trials). Note that these matrices are not standard confusion matrices, but they are closely related in that they express the pairwise difficulty (confusion) of each pair of objects. Objects in Fig 2-4A have been re-ordered based on a hierarchical clustering of human error patterns to highlight structure in the matrix. Since difficulty is measured via unbiased performance in discriminating pairs of objects, matrices are symmetric by construction. We noted high qualitative

similarities in these patterns of performance (Fig. 2-4A, compare pooled human and pooled monkey). For example, (camel, dog) and (tank, truck) are two examples of object pairs that are often confused by humans and monkeys alike.

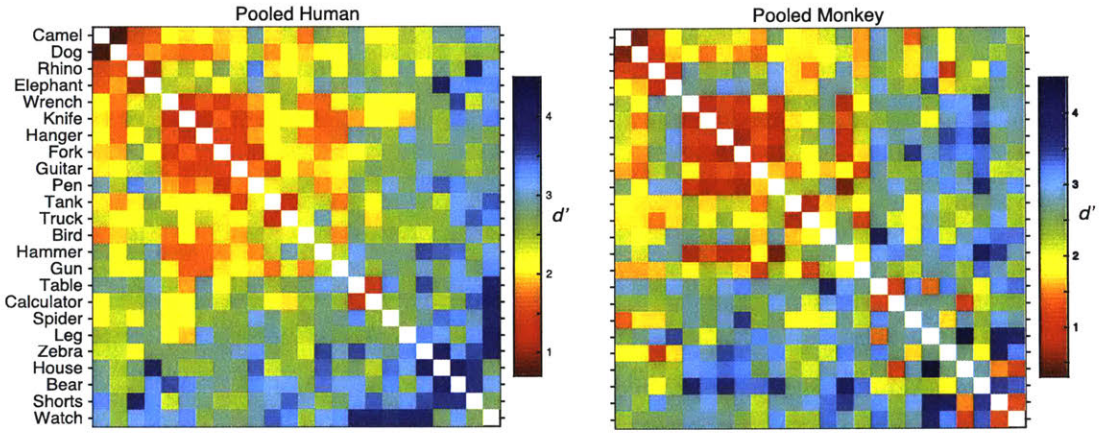

Figure 2-4: Pattern of behavioral performances for the pooled human and pooled monkey. Each 24x24 matrix summarizes confusions of all 2-way tasks: the color of bin (i,j) indicates the unbiased performance ($d\tilde{O}$) of the binary recognition task with objects i,j. Objects have been re-ordered based on a hierarchical clustering of human confusion patterns to highlight structure in the matrix. We observe qualitative similarities in the confusion patterns. For example, (camel, dog) and (tank, truck) are two often confused object pairs in both monkeys and humans.

To quantify the similarity of these patterns of behavioral performance, we took the "pooled human" pattern as the gold standard. We first compared the pooled monkey pattern by computing the noise-adjusted correlation of the 276 d' values (termed "human-consistency" see Methods). We found this number to be 0.78 ± 0.007 (Figure 2-4B, top panel). Using the same human gold standard and the same methods, we also computed the consistency of each monkey, a low-level pixel representation (Figure 2-5A, bottom panel), and computer vision representations (Figure 2-5B). Both monkeys show patterns of confusions that are highly correlated with "pooled human" confusion patterns (Monkey M: 0.80 ± 0.009; Monkey Z: 0.66 ± 0.005). Importantly, low-level visual representations do not share these patterns of behavioral performance (pixels: 0.37 ± 0.003; V1+: 0.52 ± 0.004). Of the two high-performing computer vision image representations, we found that from the top layer of a state-of-the-art deep

convolutional neural network model optimized for invariant object recognition performance was highly predictive of human confusion patterns (CNN2013: 0.86 ± 0.006), in contrast to an earlier, alternative model of the ventral stream (HMAX: 0.55 ± 0.004).

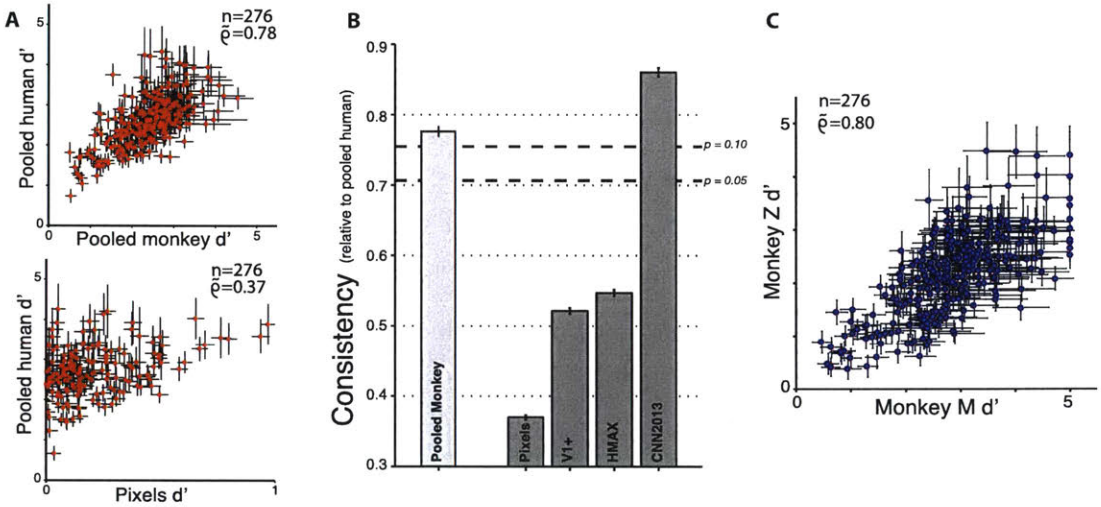

Figure 2-5: A) Comparison of d' estimates of all 276 tasks (mean $\pm$ SE as estimated by bootstrap, 100 resamples) of the pooled human with that of the pooled monkey (top panel), and a low-level pixel representation (bottom panel). B) Quantification of consistency as noise-adjusted correlation of d' vectors. The pooled monkey shows patterns of confusions that are highly correlated with pooled human subject confusion patterns ($\tilde{\rho} = 0.78$). Importantly, low-level visual representations do not share these confusion patterns (pixels: 0.37; V1+: 0.52). Furthermore, a state-of-the-art deep convolutional neural network representation was highly predictive of human confusion patterns (CNN2013: 0.86), in contrast to an alternative model of the ventral stream (HMAX: 0.55). The dashed lines indicate thresholds at $p = 0.1, 0.05$ confidence for consistency to the gold standard pooled human, estimated from pairs of individual human subjects. C) Comparison of d' estimates of all 276 tasks (mean $\pm$ SE as estimated by bootstrap, 100 resamples) between the two monkeys.

2.2.3 Natural subject-to-subject variation

While both monkeys' patterns of behavioral performance were highly consistent with the pooled human pattern, they were not perfectly correlated (i.e. the consistency value is not 1.0). We asked if this reflected a true species difference between human and monkey behavior or whether it might be explained by within-species subject

variability. To do so, 16, 16 and 12 separate MTurk subjects were recruited to characterize individual human subject behavior on three groups of eight objects. Firstly, we observed that there is significant inter-subject variability in the tested human population; the median consistency of behavioral patterns between pairs of individual humans subjects is only 0.76. Consequently, individual human subjects are on average not perfectly correlated with the "pooled human". Figure 2-5 shows both absolute performance (percent correct) and consistency (relative to the "pooled human") of individual human subjects, individual monkeys, and low-level machine features on the three tested groups of eight objects. Note that these data account for only 30% (3*28/276) of all tasks presented in Figure 2-4. The solid and dashed line indicate the mean $\pm$ SD performance/consistency of individual human subject population; for the three groups of eight objects, the mean consistency ($\pm$ SD) of individual human subjects were 0.74 ± 0.18, 0.77 ± 0.11, 0.68 ± 0.18. Importantly, this variability is sufficiently small to reject some representations. Indeed, low-level visual representations that model the retina and primary visual cortex fall outside the distribution of consistency over human subjects for two out of three groups of objects (Figure 2-5; V1+: $p = 0.30, 0.03, 0.03$; pixels: $p = 0.17, 0.02, 0.03$; Fisher's exact test for the three groups of objects respectively). However, both monkeys remain statistically indistinguishable from individual human subjects in their consistency to the "pooled human" (monkey M: $p = 0.4, 0.21, 0.22$; monkey Z: $p = 0.16, 0.22, 0.27$ for the three groups; Fisher's exact test). Additionally, the CNN2013 model could not be falsified in any of the three groups of objects ($p = 0.38, 0.35, 0.36$ for the three groups of objects respectively) while HMAX was rejected in one of the three groups ($p = 0.26, 0.07, < 0.01$ respectively).

We next asked whether inter-subject variability might also explain the imperfect consistency of the pooled monkey relative to the "pooled human" (see Figure 2-4C). To account for the small sample size of monkeys ($n = 2$), we randomly sampled pairs of individual human subjects, and measured the consistency relative to the "pooled human" of their pooled behavioral data. This process was repeated 50 times for each

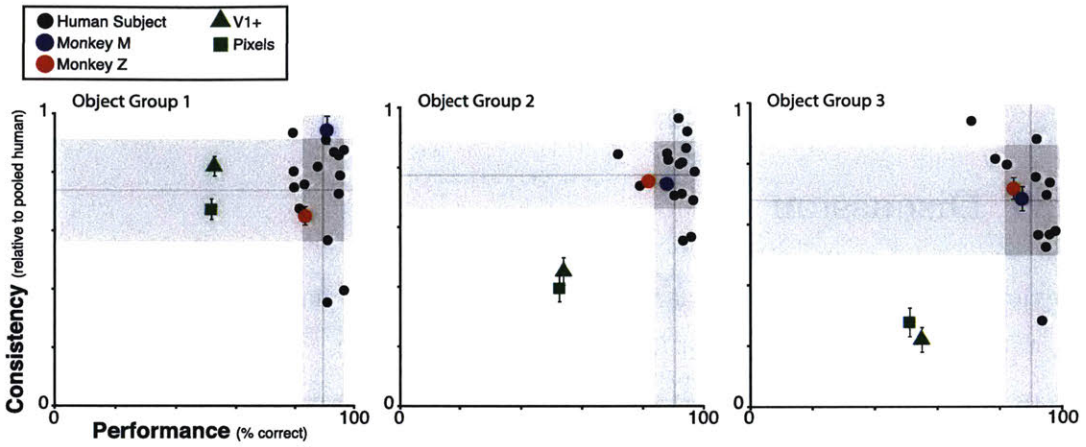

Figure 2-6: Accounting for inter-subject variability. For each of three groups of eight objects, the absolute performance and consistency of individual human subjects, individual monkeys, and machine features are shown. Error bars for consistency relative to pooled human (mean ± SD) are shown for individual monkeys and machine features for each group (error bars for monkeys are not visible in object group 2 due to small variability). The shaded grey areas indicate the distribution of performance/consistency over individual human subjects (mean ± SD). There is significant inter-subject variability: individual human subjects are on average not perfectly correlated with the pooled human (average consistency 0.74, 0.77, 0.68 for the three groups). As a result, monkeys are statistically indistinguishable from individual human subjects in their correlation to the human pool. In contrast, low-level visual representations were falsified on both performance and consistency grounds for two out of three groups of objects.

of the three groups of eight objects to obtain a distribution of consistency of $n = 2$ pooled human subjects. Figure 2-4C shows the $p = 0.1$ and $p = 0.05$ confidence thresholds of this distribution (dashed lines). The pooled monkey cannot be rejected relative to this distribution, i.e. the pooled monkey's patterns of performance are statistically indistinguishable from patterns of similarly sampled pools of human subjects.

We also estimated biases in object confusion patterns using the criterion index c (see Methods). We found this measure was significantly less replicable across subjects: the median consistency of bias (c) between pairs of individual humans subjects was 0.40, compared to 0.76 for consistency of unbiased performance (d'), suggesting that

biases are significantly less meaningful behavioral signatures on which to compare humans and monkeys.

2.3 Discussion

Previous work has revealed quantitative similarity of humans and macaque monkey behavior in low-level visual behaviors [De Valois et al., 1974a, De Valois et al., 1974b, Vogels and Orban, 1990, Vazquez et al., 2000, Kiorpes et al., 2008, Gagin et al., 2014], suggesting that an understanding of the neural mechanisms underlying those tasks in macaques will directly translate to humans. However, such correspondences for high-level visual behaviors such as view-invariant object recognition have not been demonstrated. While many studies have shown that monkeys can learn to perform tasks based on object shape [Mishkin et al., 1983, Minamimoto et al., 2010, Okamura et al., 2014], this is taken by some as evidence of the powerful learning abilities of monkeys in experimenter-chosen tasks, rather than a tight correspondence with humans in their behavioral patterns in object discrimination abilities. In this study, we systematically compared the basic-level core object recognition behavior of two rhesus macaque monkeys with that of human subjects. Our results show that monkeys not only match human performance, but show a pattern of object confusion that is highly correlated with "pooled human" confusion patterns, and that individual monkey subjects are statistically indistinguishable from the population of individual human subjects. Importantly, these shared patterns of basic-level object confusions are not shared with low-level visual representations (pixels, V1+).

We characterized average human population and individual human subject behavior using high-throughput online psychophysics on Amazon's Mechanical Turk system. This method allowed us to efficiently gather datasets of otherwise unattainable sizes, and has previously been validated by comparing results obtained from online and in-lab psychophysical experiments [Crump et al., 2013]. In particular, patterns of behavioral performance on object recognition tasks from in-lab and on-

line subjects were equally reliable and virtually identical [Majaj et al., 2015]. While we did not impose experimental constraints on subjects' acuity, and we can only infer likely gaze position, the observed high performance and highly reliable confusion patterns suggest that this sacrifice in exact experimental control is a good trade off for the very large number of tasks (276) that could be tested.

We characterized monkey object recognition behavior from two subjects using two different effector systems. This modest sample size is typical for systems neuroscience experiments, given the cost and difficulty of monkey psychophysics. As a result, it is unclear whether the differences observed between monkeys (consistency between monkeys: 0.80) reflect true inter-subject variability, or are due to differences in effector system. Monkey Z's overall performance (83.4%) was lower than monkey M's (89.2%) and, for an equal number of confusions, confusion patterns from monkey Z were significantly less reliable than those from monkey M ($p < 0.001$, two-sample t-test). These differences suggest additional variance ("noise") in monkey Z's behavioral data, potentially due to less gaze control than monkey Z, that may partly account for the differences in behavioral patterns between monkeys. However, this additional behavioral variance did not significantly impact the result; each monkey subject was highly correlated with the human pool, and statistically indistinguishable from individual humans.

Additionally, it is possible that we failed to observe a significant difference between monkeys and humans due to a lack of statistical power from a sample of just two monkeys. In principle, one cannot prove that there is absolutely no difference between monkeys and humans, as ever-increasing power would be required for the testing of an ever-smaller proposed difference. Here, we showed that our behavioral tests do have sufficient power to falsify other models (pixels, V1+) as matching human core object recognition behavior, but failed to falsify monkeys as a model of that domain of human behavior. Testing additional monkeys on this behavioral domain, or additional behavioral tests beyond this domain may, in principle, reveal differences

between monkey and human object recognition behavior.

We argue that the observed results are not due to overtraining of animals. Monkeys were trained using a standard operant conditioning paradigm to report object identity in visual images. Objects were novel to monkeys prior to behavioral training. When presented with these novel objects, monkeys were able to reach high-level performance relatively quickly (see Figure 2-3A-C). Furthermore, by sampling from a large pool of images, we were able to ensure that monkeys were exposed to any given image at most once per behavioral session on average. Finally, we collected behavioral data on a set of held-out images (test set, 100 images/object) after monkeys were fully trained to criterion on all tasks. Importantly, both monkeys successfully generalized to these new images of previously learned objects; performance on the very first trial of a new image was high for both monkeys (Monkey M: 88%, Monkey Z: 85%), and the first-trial performance was not predicted by the similarity of test images to previously seen training images (see Figure 2-3D). As a result, the measured patterns of behavioral performance reflect the monkeys' ability to discriminate between pairs of basic-level objects, rather than memorized or overtrained behavior. Importantly, humans, while not explicitly trained on these images, likely get significant prior experience with similar objects over the course of their lifetimes. We observed that individual humans perform at a high initial performance, and exhibit no change in performance as a function of (unsupervised) exposure to objects (see Figure 2-3C), suggesting that humans are already well "trained" on these tasks. In sum, while it is impossible to guarantee or expect that humans and monkeys have identical lifetime experience, we find that, once monkeys were trained, further experience has little to no effect on the patterns of behavioral performance of either species. We note that this does not imply that monkeys and humans learn new objects at a similar rate, only that their steady state patterns of behavior are highly comparable.

Object categories consisted of basic-level objects with a single object instance (a single 3D model) per category. Consequently, our results do not speak to monkeys'

ability to generalize across multiple object instances within semantic categories, but are instead constrained to judgments of visual similarity of 3D objects. Species differences at the category level are possible. Similarly, past studies have debated about differences in the specificity of the "face-inversion effect" between macaque monkeys and chimpanzees/humans [Bruce, 1982, Vermeire and Hamilton, 1998, Parr, 2011]. Our results do not rule out the possibility of such species differences for subordinate level object recognition behaviors. Future work with semantic object categories or subordinate-level object recognition tasks would thus be important for discerning the limits of the species comparison over all of object recognition behavior.

The observed similarities in monkey and human object recognition behavior are consistent with comparative functional imaging of macaque and human brains that reveal strong species homologies of visual cortical areas [Orban et al., 2004], particularly object-selective regions, based on activity correlations [Mantini et al., 2012] and connectivity [Miranda-Dominguez et al., 2014]. While strict anatomical homologies may be imperfect due to evolution-driven reorganization, functional measures reveal a near-perfect conservation of the ventral visual stream, a hierarchy of visual cortical areas thought to directly underlie object recognition behavior, between macaque monkey and human [Mantini et al., 2012]. In particular, the neuronal representations of object categories in the end-stage of the ventral stream are matched between monkey and human [Kriegeskorte et al., 2008]. Taken together, the anatomical, physiological and behavioral similarities between monkeys and humans are consistent with the hypothesis that monkeys and humans share similar neural representations underlying the visual perception of basic-level objects.

Recent advances in machine learning have uncovered high-performing representations for object recognition using deep convolutional neural network models [LeCun et al., 2015]. Interestingly, these computational models rival the primate brain for core object recognition behavior [Cadieu et al., 2014] and accurately predict neural responses of high-level visual cortical representations of monkey [Yamins et al., 2014]

and humans [Khaligh-Razavi and Kriegeskorte, 2014]. Here, we report that monkey and human behavioral patterns were well predicted by a state-of-the-art deep convolutional neural network model (CNN2013), in contrast to alternative models of the ventral stream (HMAX) and low-level control models (pixels, V1+). Taken together, these results suggest that current high-performing deep convolutional neural network models may accurately capture the shared representation that directly underlies core basic-level object recognition in both humans and monkeys.

To conclude, systematic comparisons of animal model and human behavior are, to date, largely lacking in the domain of invariant visual object recognition. Here, we investigated whether this behavior is quantitatively comparable across rhesus monkeys and humans. Our results show that monkeys and humans are statistically indistinguishable on a large battery of basic-level visual object recognition tasks, suggesting that rhesus monkeys and humans may share a neural "shape" representation that directly underlies object perception, and supporting the use of the monkey as a closely matched model system for studying ventral stream visual processing and object representation in humans.

2.4 Methods

2.4.1 Visual images

We examined "basic-level" object recognition behavior by generating images of a set of 64 objects that we previously found to be highly reliably labeled by independent human subjects, based on the definition proposed by [Rosch et al., 1976]. From this set, three groups of eight objects were sampled for this study; the selection of these 24 objects was random, but biased towards groups of objects which exhibited reliable confusion patterns in humans (see Fig. 1 for a full list of those 24 basic-level objects). To enforce true object recognition behavior (rather than image matching), several thousand naturalistic images, each with one foreground object, were generated

by rendering a 3D model of each object with randomly-chosen viewing parameters (2D position, 3D rotation and viewing distance) and placing that foreground object view onto a randomly-chosen, natural image background. To do this, each object was first assigned a canonical position (center of gaze), scale ($\sim 2°$) and pose, and then its viewing parameters were randomly sampled uniformly from the following ranges for object translation ($[-3, 3]°$ in both h and v), rotation ($[-180, 180]°$ in all three axes) and scale ([x0.7, x1.7]. Backgrounds images were sampled randomly from 3D HDR images of indoor and outdoor scenes obtained from Dosch Design (www.doschdesign.com). As a result, these images require any visual recognition system (human, animal or model) to tackle the "invariance problem," the computational crux of object recognition, as it is highly challenging for low-level visual representations [Ullman and Humphreys, 1996, Pinto et al., 2008]. Using this procedure, we generated 2400 "test" images (100 images per object) at 1024x1024 pixel resolution with 256-level grayscale and with square apertures for human psychophysics, monkey psychophysics and model evaluation. A separate set of 4800 "training" images (200 images per object) were generated with the same procedure with circular apertures to initially train the monkeys. Figure 2-1 shows example test images for each of the 24 basic-level objects.

To quantify the overlap between training and test image sets, we computed the pixel Euclidean distance of each test image to the nearest training image of the same object. For this analysis, training and test images were re-rendered on gray backgrounds to measure background-independent distances, and resized to 64×64 pixel resolution. The resulting distance distribution was compared to that computed from simulated "ideal" generalization conditions. We rendered six different surrogate training image sets, each with identical generative parameters to the background-less training images except for one of the six viewpoint parameters held at its mean value. These sparsely sampled training images simulated six different experimental conditions wherein subjects would not have been exposed to variations in one parameter during the training stage, but later tested on full variation images. From these newly

generated training images, we computed the corresponding "ideal" pixel Euclidean distances of each test image to the nearest training image of the same object. We found that the distribution of background-independent distances of actual test images from actual training image sets was only slightly less than the corresponding distribution across the simulated ideal generalization conditions (3.4% and 3.7% increase in median and maximum distances respectively for the simulated conditions; see Figure 2-3D, bottom panel). This suggests that our 4800 training images did not sample the image space too "densely", but rather of comparable sparsity as if we had entirely held back particular types of viewpoint variations.

2.4.2 Human Behavior

All human behavioral data presented here were collected from 638 human subjects on Amazon's Mechanical Turk (MTurk) performing 276 interleaved, basic-level, invariant, core object recognition tasks. Each task consisted of a binary discrimination between pairs of objects from the 24 objects considered. Subjects were instructed to report the identity of the foreground object in each presented image, from two given choices. Because those two choices were provided after the test image and all 276 tasks were randomly interleaved (trial-by-trial), subjects could not deploy feature attentional strategies specific to each task to process the test images. Each trial initiated with a central black point for 500ms, followed by 100ms presentation of a test image. The test image contained one foreground object presented under high variation in viewing parameters and overlaid on a random background, as described in the Visual Images section (above). Immediately after extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were shown to the left and right. One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. After mouse clicking on one of the choice images, the subject was given another fixation point before the next stimulus appeared. No feedback was given; subjects were never explicitly trained on the tasks. Under assumptions of typical computer ergonomics, we estimate that images were

presented at $6 - 8°$ of visual angle in size, and response images were presented at $6 - 8°$ of eccentricity.

The online Mechanical Turk platform enables efficient collection of reliable, large-scale psychophysical data, and has been validated by comparing results obtained from online and in-lab psychophysical experiments [Crump et al., 2013]. In particular, a previous study from our group directly compared the patterns of behavioral performance on invariant object recognition tasks of in-lab and online subjects. In brief, human subjects were tested in a controlled in-lab setting on eight-alternative forced choice core object recognition tasks, at both basic and subordinate levels. A total of 10, 15, and 22 subjects each performed 600 trials of basic-level object categorization, car identification and face identification tasks respectively. Pooling trials from all subjects, the in-lab human behavioral data was highly reliable ($\tilde{\rho} = 0.95 \pm 0.024$; median±SE). A separate set of 104 human subjects from Amazon's Mechanical Turk performed trials of the same tasks, resulting in similarly reliable pooled online human data ($\tilde{\rho} = 0.97 \pm 0.023$; median±SE). Accounting for noise, the behavioral patterns of in-lab and online human subjects were virtually identical ($\tilde{\rho} = 0.98$, see Analysis section) [Majaj et al., 2015], supporting the use of MTurk for characterizing human core object recognition behaviors. Following the methods of that prior work, we here did not perform eye tracking of online human subjects to measure or control their gaze. Instead, subjects were cued to the location of image presentation with a fixation cue. Subjects detected as obviously guessing were banned from further experiments and the corresponding data were excluded from further analyses (less than 1% of subjects were eliminated). To do this, we quantified this guessing behavior using a choice entropy metric, which measured how well a subject's current trial response was predicted by the previous trial response. For all remaining subjects, we did not observe any significant differences in performance between the first and last halves of behavioral trials ($p = 0.49$, t-test), suggesting subjects did not undergo substantial learning. Overall, subjects achieved high performance on all behavioral tasks ($88.35\% \pm 5.6\%$, mean ± SD, n = 276 tasks).

Most of the human psychophysical subjects were used to characterize "pooled human" behavior. Specifically, data from 605 MTurk subjects each performing a relatively small number of trials (mean of 114 trials/subject) were aggregated to obtain a highly reliable estimate of "pooled human" object recognition performance on each task. Each subject only performed a subset of the tasks (mean of 67 tasks/subject). All trials of all of these subjects (69,000 trials in total, 250 trials/task) were pooled together to characterize "pooled human" behavior.

A separate set of human subjects was used to characterize the variability in individual human subject behavior. Specifically, these MTurk subjects performed a relatively large number of trials of binary object recognition tasks over groups of only eight of the 24 objects (note: 8 objects generates 28 unique binary tasks). Each individual human subject performed trials for all of those 28 tasks. To ensure that sufficiently many trials were collected from each individual subject for reliable measurement of his or her pattern of behavioral performance, we used a pre-determined criterion for reliability, defined as split-half internal consistency (see Analysis section) significantly greater than 0.5 ($p < 0.05$, one-tailed t-test). We then attempted to repeatedly recruit each subject until his or her total pool of data reached this reliability criterion. Of 80 unique subjects that performed this task, 33 were succcssfully re-recruited a sufficient number of times to reach the reliability criterion on at least one group of objects (five of these subjects performed tasks in two different groups, and three subjects performed in all three groups). In total, we collected data from 16, 16, and 12 subjects for each of three groups of eight objects (mean of 3,003 trials/subject within each group).

Humans, while not explicitly trained on these images, likely get extensive experience with similar objects over the course of their lifetime. To investigate the effect of experimental experience on behavior, we measured the performance of individual human subjects as a function of the number of presentations per object. To allow di-

rect comparison with monkeys, this analysis was constrained to the 16 objects in the second and third groups for which corresponding monkey "training" data were also available. Figure 2-3C shows the relative performance, quantified as a Turing ratio ($\frac{d'\text{human subject}}{d'\text{human pool}}$, one-versus-all d'), of individual humans subjects with sufficient longitudinal data on these tasks (defined as >150 trials/object, 15 unique human subjects). We observe that individual humans perform at a high initial performance, and exhibit no change in performance as a function of (unsupervised) experience with the objects, suggesting that humans are already well trained on these tasks.

2.4.3 Monkey Training and Behavior

Monkey behavioral data on the exact same object recognition tasks were collected from two adult male rhesus macaque monkeys (*Macaca mulatta*) weighing 6 kg (monkey M) and 12 kg (monkey Z). All procedures were performed in compliance with National Institutes of Health guidelines and the standards of the MIT Committee on Animal Care and the American Physiological Society. To ensure that our behavioral tests were tapping a sensory representation (i.e. did not depend on the reporting effector), we tested one monkey (M) using saccade reports (gaze tracking) and the other monkey (Z) using reaching reports (touch screen).

Monkey M: Prior to behavioral training, a surgery using sterile technique was performed under general anesthesia to implant a titanium head post to the skull. Following head-post implant surgery, Monkey M was trained on a match-to-sample paradigm under head fixation and using gaze as the reporting effector. Eye position was monitored by tracking the position of the pupil using a camera-based system (SR Research Eyelink II). Images were presented on a 24" LCD monitor (1920 x 1080 at 60 Hz; Acer GD235HZ) positioned 42.5 cm in front of the animal. At the start of each training session, the subject performed an eye-tracking calibration task by saccading to a range of spatial targets and maintaining fixation for 800 ms. Calibration was repeated if drift was noticed over the course of the session. Figure 2-2B illustrates the behavioral paradigm. Each trial was initiated when the monkey acquired and held

gaze fixation on a central fixation point for 200ms, after which a test image appeared at the center of gaze for 100ms. Trials were aborted if gaze was not held within $\pm 2^\circ$. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right (each centered at 6° of eccentricity along the horizontal meridian; see Fig. 1B). One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1500ms, and indicated its final choice by holding fixation over the selected image for 700ms. During initial training, the monkey typically visually explored both objects before making a selection, but quickly transitioned to selecting its choice covertly in that it often directed its first saccade in the direction of the final choice.

Monkey Z performed the same task using a touch screen — other than the differences noted below, the task was identical to Monkey M. Monkey Z underwent no surgical procedures, and was instead seated head-free in front of a 15" LCD touchscreen (1024 x 768 at 60 Hz, ELO Touch 1537L) at a distance of 34.2cm. The subject interacted with the task by touching the screen with his left hand through an opening in the primate chair. Monkey Z initiated each trial by touching a fixation point 4° below the center of the screen for 250 ms, which triggered the presentation of the test image at the center of the screen (i.e. this ensured that the hand and finger rising from below did not occlude any of the test image). After the appearance of the choice images, the monkey indicated its choice by touching the selected image. Gaze was not controlled or measured in Monkey Z, but we instead assumed that touch point acquisition would correspond to gaze being directed at the screen. Because the test image screen location was fixed over trials and the test image content was required for successful reward, we assumed that the animal's gaze would be reliably directed at the center of each test image. This assumption is supported by the finding that Monkey Z showed a very similar pattern of performance as Monkey M (Fig. 3D).

The images were sized so that they subtended $6 \times 6°$ for each monkey. Realtime experiments for all monkey psychophysics were controlled by open-source software (MWorks Project http://mworks-project.org/). Animals were rewarded with small juice rewards for successfully completing each trial, and received time-outs of 1.5 to 2.5 seconds for incorrect choices. Animals were trained to work with each group of eight objects to criterion, defined as a stable pattern of behavioral performance over at least four behavioral sessions, before moving on to the next group. For the first group, both monkeys were exposed to images with gradually increasing variation in viewing parameters (pose, position and viewing distance) over several behavioral sessions. For each subsequent group of eight objects, animals were immediately exposed to full variation images, and reached high performance in 1000-2000 image presentations per object ($\tilde{}10 - 15$ behavioral sessions). Figure 2-3 shows the monkeys' performance relative to the human pool, quantified as a Turing ratio ($\frac{d'\text{monkey}}{d'\text{human pool}}$, one-versus-all d'), for each of these 16 objects. When presented with these novel objects, monkeys were able to reach high-level performance relatively quickly (see Figure 2-3A,B). Following training of all binary tasks in each group of eight objects, animals were trained to criterion on the remaining pairs of objects. Once animals reached criterion on all 276 possible binary object recognition tasks, complete behavioral data was collected in a fully-interleaved manner, first using training images and subsequently switching to held-out test images. Importantly, monkeys immediately generalized to new images of previously learned objects, with comparable high performance on the very first trial of a new image for both monkeys (Monkey M: 88%, Monkey Z: 85%). Furthermore, the monkeys' performance on the first trial of novel test images was not dependent on the test image's similarity to previously seen training images (see Methods - Visual Images section). We observed no significant negative correlation between first-trial performance of test images and their background-independent distance to the nearest training images ($r = 0.036$, $p = 0.07$ and $r = 0.010$, $p = 0.63$ for monkey M and Z respectively), as shown in the top panel of Figure 2-3D (mean $\pm$ SE, pooling both monkeys). Subsequent exposures to these test images did not further increase behavioral performance (Figure 2-3D, zero

distance marker). Taken together, this suggests that monkeys did not rely simply on the similarity to previously seen images. Furthermore, the object confusion patterns were found to be largely independent of the image set; the consistency (computed as a noise-adjusted correlation, see Analysis section) between confusion patterns of the training and test image sets was $\tilde{\rho} = 0.9566 \pm 0.0253$ and 0.9489 ± 0.0157 (mean $\pm$ SD, for monkey M and Z respectively). Thus, complete behavioral data collected in a fully interleaved manner from both images sets were pooled. A total of 106,844 trials were collected from both monkeys (51,096 from Monkey M and 55,748 from Monkey Z) and used for the analyses below.

2.4.4 Machine Behavior

We tested different machine systems on our 276 tasks by computing each machine's feature population output for each of our images, and using trained classifiers to make behavioral choices based on the test images.

Low-level visual representations of pixel and V1+ [Pinto et al., 2008] features were used as control models. These features approximate the representations of the retina and primary visual cortex respectively. High-performing feature representations from state-of-the-art computer vision models were also tested for comparison. HMAX ([Riesenhuber and Poggio, 1999, Serre et al., 2007] is a model inspired by the tolerance and selectivity properties of the ventral visual stream. We used the publicly available FHLib implementation [Mutch and Lowe, 2008]. We trained the model on 5760 synthetic images of 3D objects drawn from eight natural categories (animals, boats, cars, chairs, faces, fruits, planes and tables; see [Yamins et al., 2014] that did not overlap with the 24 objects used in this study. CNN2013 refers to a model based on a state-of-the-art deep convolutional neural network model [Zeiler and Fergus, 2014]. Using an architecture and learning parameters based on [Zeiler and Fergus, 2014], we trained a deep convolutional neural network for 15 epochs on images drawn from the ImageNet 2013 challenge set, adjusting the learning rate in accordance with

the heuristics described in this publication. We used a publicly available implementation [Wan et al., 2013], itself based on CudaConvnet [Krizhevsky et al., 2012] that allowed for dropout regularization. Training the model took two weeks on a Tesla K40 GPU, provided by NVIDIA.

For each machine representation, features were extracted from the same images that were presented to humans and monkeys. As with humans and monkeys, each machine representation was tested on the same 276 interleaved binary object recognition tasks. For each machine feature representation, performance on each binary task was measured using two-fold cross validation using a maximum correlation classifier, repeated 50 times over permutations of classifier training and testing data partitions.

2.4.5 Analysis

Behavioral metric and behavioral consistency

For each of the 276 binary object recognition tasks, we estimated an unbiased measure of performance using a sensitivity index d' [Macmillan, 1993]: $d' = Z(\text{HitRate}) - Z(\text{FalseAlarmRate})$, where $Z(.)$ is the inverse of the cumulative Gaussian distribution. Hit rates and false alarm rates were computed at the resolution of objects, pooling over all trials and images of each object. All d' estimates were constrained to a range of [0,5]. Bias was estimated using a criterion index c [Macmillan, 1993]: $c = 0.5 * (Z(\text{HitRate}) + Z(\text{FalseAlarmRate}))$. We here refer to the 276-dimensional vector of d' values over all binary object recognition tasks as the "pattern of behavioral performance" (b).

The reliability (a.k.a. internal consistency) of behavioral data was computed as the Spearman correlation between patterns of behavioral performance patterns computed on separate halves of the data (random split-halves of trials); this process was repeated across 100 draws. Since this estimate is measured using only half of the data, the Spearman-Brown prediction formula [Brown, 1910, Spearman, 1910] was

applied to allow for comparisons to correlations measured using all trials.

Consistency between different behavioral patterns $b1, b2$ was then computed as a noise-adjusted rank correlation between patterns of behavioral performances (d' or c vectors):

$$\widetilde{\rho} = \frac{\rho_{b1,b2}}{\sqrt{\rho_{b1,b1} \times \rho_{b2,b2}}}$$

where $\rho_{b1,b2}$ is the raw Spearman rank correlation, and $\rho_{b1,b1}$, $\rho_{b2,b2}$ are the Spearman-Brown corrected internal consistency estimates of each behavioral pattern. Our rationale for using a noise-adjusted correlation measure for consistency is to account for variance in the behavioral patterns that arises from "noise", i.e. variability that is not replicable by stimulus object identity [DiCarlo and Johnson, 1999, Johnson et al., 2002]. We obtained a distribution of consistency values using the 100 resampled estimates of internal consistencies of each behavioral pattern (i.e. from the 100 random draws of split-halves of trials of b1, b2).

2.5 Acknowledgements

This research was performed in collaboration with Kailyn Schmidt and James J. DiCarlo.

Chapter 3

Comparison of visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks

A primary neuroscience goal is to uncover neuron-level mechanistic models that quantitatively explain this behavior by predicting primate performance for each and every image. Recently, specific feed-forward deep convolutional artificial neural networks (ANNs) models have dramatically advanced our quantitative understanding of the neural mechanisms underlying primate core object recognition. In this work[1], we tested the limits of those ANNs by systematically comparing the behavioral responses of these models with the behavioral responses of humans and monkeys, at the resolution of individual images. Using these high-resolution metrics, we found that all tested ANN models significantly diverged from primate behavior. Going forward, these high-resolution, large-scale primate behavioral benchmarks could serve as direct guides for discovering better ANN models of the primate visual system.

[1]The contents of this chapter are adapted from a journal article in preparation [Rajalingham et al., 2018].

3.1 Introduction

Primates—both human and non-human—can typically recognize objects in visual images at a glance, even in the face of naturally occurring identity-preserving transformations such as changes in viewpoint. This view-invariant visual object recognition ability is thought to be supported primarily by the primate ventral visual stream [DiCarlo et al., 2012]. A primary neuroscience goal is to construct computational models that quantitatively explain the neural mechanisms underlying this ability. That is, our goal is to discover artificial neural networks (ANNs) that accurately predict neuronal firing rate responses at all levels of the ventral stream and its behavioral output. To this end, specific models within a large family of deep, convolutional neural networks (DCNNs), optimized by supervised training on large-scale category-labeled image-sets (ImageNet) to match human-level categorization performance [Krizhevsky et al., 2012, LeCun et al., 2015], have been put forth as the leading ANN models of the ventral stream [Yamins and DiCarlo, 2016]. We refer to this sub-family as DCNN_{IC} models (IC to denote ImageNet-categorization pre-training), so as to distinguish them from all possible models in the DCNN family, and more broadly, from the super-family of all ANNs. To date, it has been shown that DCNN_{IC} models display internal feature representations similar to neuronal representations along the primate ventral visual stream [Yamins et al., 2014, Cadieu et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014], and they exhibit behavioral patterns similar to the behavioral patterns of pairwise object confusions of primates [Rajalingham et al., 2015]. Thus, DCNN_{IC} models may provide a quantitative account of the neural mechanisms underlying primate core object recognition behavior.

However, several studies have shown that DCNN_{IC} models can diverge drastically from humans in object recognition behavior, especially with regards to particular images optimized to be adversarial to these networks [Goodfellow et al., 2014, Nguyen et al., 2015]. Related work has shown that specific image distortions are disproportionately challenging to current DCNNs, as compared to humans [RichardWebster

et al., 2016, Dodge and Karam, 2017, Geirhos et al., 2017, Hosseini et al., 2017]. Such image-specific failures of the current ANN models would likely not be captured by "object-level" behavioral metrics (e.g. the pattern of pairwise object confusions mentioned above) that are computed by pooling over hundreds of images and thus are not sensitive to variation in difficulty across images of the same object. To overcome this limitation of prior work, we here aimed to use scalable behavioral testing methods to precisely characterize primate behavior at the resolution of individual images and to directly compare leading DCNN models to primates over the domain of core object recognition behavior at this high resolution.

We focused on core invariant object recognition—the ability to identify objects in visual images in the central visual field during a single, natural viewing fixation [DiCarlo et al., 2012]. We further restricted our behavioral domain to basic-level object discriminations, as defined previously [Rosch et al., 1976]. Within this domain, we collected large-scale, high-resolution measurements of human and monkey behavior (over a million behavioral trials) using high-throughput psychophysical techniques—including a novel home-cage behavioral system for monkeys. These data enabled us to systematically compare all systems at progressively higher resolution. At lower resolutions, we replicated previous findings that humans, monkeys, and DCNN_{IC} models all share a common pattern of object-level confusion [Rajalingham et al., 2015]. However, at the higher resolution of individual images, we found that the behavior of all tested DCNN_{IC} models was significantly different from human and monkey behavior, and this model prediction failure could not be easily rescued by simple model modifications. These results show that current DCNN_{IC} models do not fully account for the image-level behavioral patterns of primates, suggesting that new ANN models are needed to more precisely capture the neural mechanisms underlying primate object vision. To this end, large-scale high-resolution behavioral benchmarks, such as those obtained here, could serve as a strong top-down constraint for efficiently discovering such models.

3.2 Results

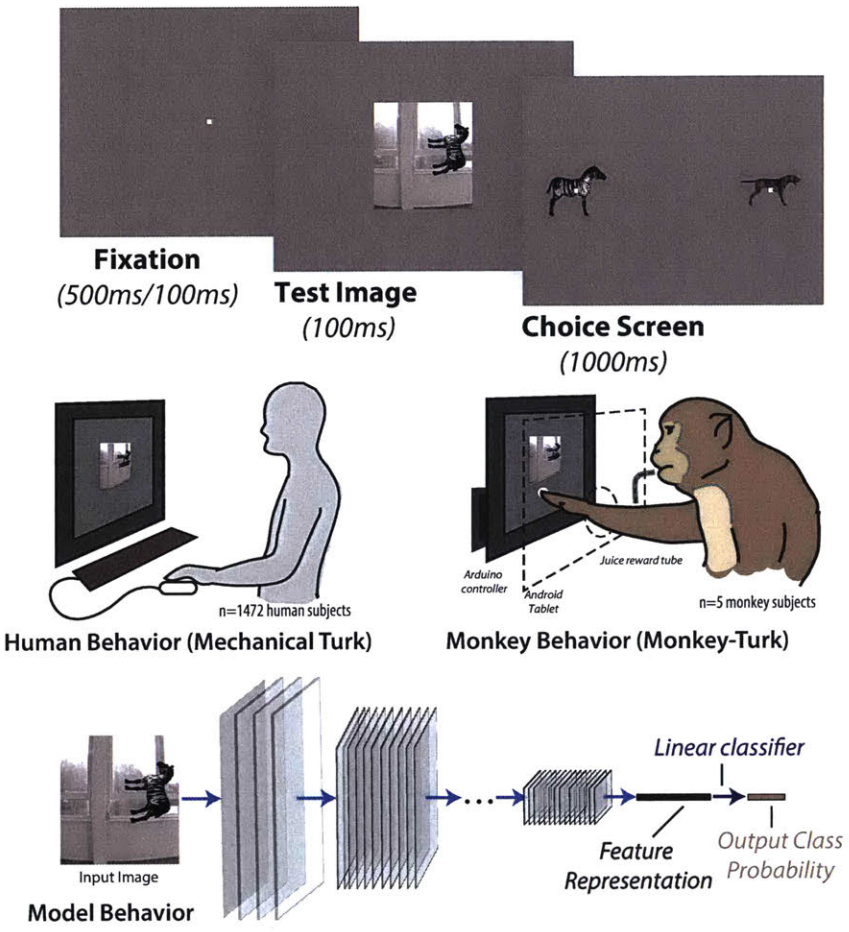

Figure 3-1: Time course of example behavioral trial (zebra versus dog) for human psychophysics. Human behavior was measured using the online Amazon MTurk platform, which enabled the rapid collection over 1 million behavioral trials from 1472 human subjects. Monkey behavior was measured using a novel custom home-cage behavioral system (MonkeyTurk), which leveraged a web-based behavioral task running on a tablet to test many monkey subjects simultaneously in their home environment. *DCNN* models were tested on the same images and tasks as those presented to humans and monkeys by extracting features from the penultimate layer of each visual system model and training back-end multi-class logistic regression classifiers.

In the present work, we systematically compared the basic level core object recognition behavior of primates and state-of-the-art artificial neural network models using a series of behavioral metrics ranging from low to high resolution within a two-alternative

forced choice match-to-sample paradigm. The behavior of each visual system, whether biological or artificial, was tested on the same 2400 images (24 objects, 100 images/object) in the same 276 interleaved binary object recognition tasks (see Figure 3-1). Each system's behavior was characterized at multiple resolutions (see Behavioral metrics and signatures in Methods) and directly compared to the corresponding behavioral metric applied on the archetypal human (defined as the average behavior of a large pool of human subjects tested; see Methods). The overarching logic of this study was that, if two visual systems are equivalent, they should produce statistically indistinguishable behavioral signatures with respect to these metrics. Specifically, our goal was to compare the behavioral signatures of visual system models with the corresponding behavioral signatures of primates.

3.2.1 Object-level behavioral comparison

We first examined the pattern of one-versus-all object-level behavior (termed "B.O1 metric") computed across all images and possible distractors. Since we tested 24 objects here, the B.O1 signature was 24 dimensional. Figure 3-2A shows the B.O1 signatures for the pooled human (pooling n=1472 human subjects), pooled monkey (pooling n=5 monkey subjects), and several DCNN_{IC} models as 24-dimensional vectors using a color scale. Each element of the vector corresponds to the system's discriminability of one object against all others that were tested (i.e. all other 23 objects). The color scales span each signature's full performance range, and warm colors indicate lower discriminability. For example, red indicates that the tested visual system found the object corresponding to that element of the vector to be very challenging to discriminate from other objects (on average over all 23 discrimination tests, and on average over all images). Figure 3-2B directly compares the B.O1 signatures computed from the behavioral output of two visual system models—a pixel model (top panel) and a DCNN_{IC} model (Inception-v3, bottom panel)—against that of the human B.O1 signature. We observe a tighter correspondence to the human behavioral signature for the DCNN_{IC} model visual system than for the baseline pixel model visual system. We quantified that similarity using a noise-adjusted correlation

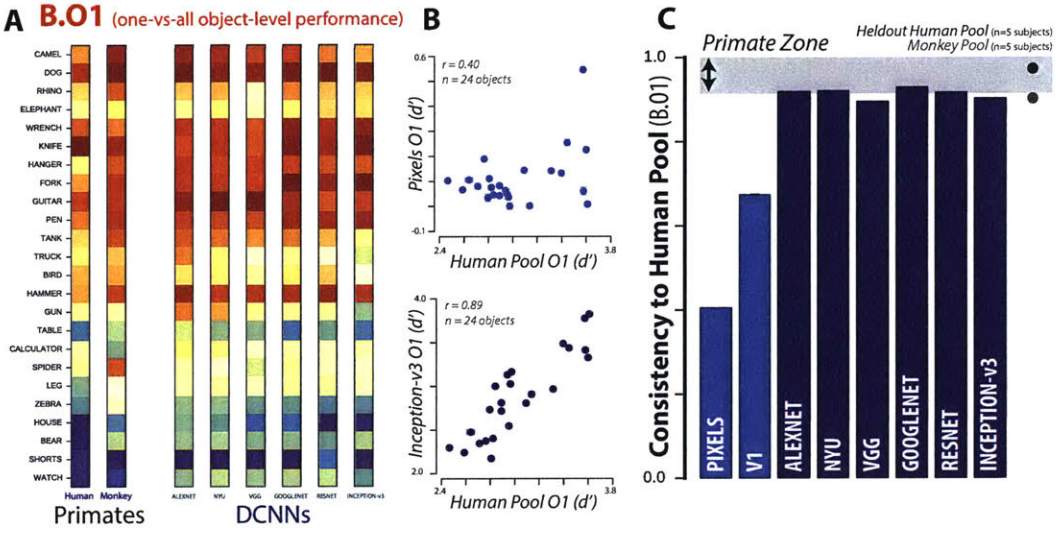

Figure 3-2: (A) One-versus-all object-level (B.O1) signatures for the pooled human (n=1472 human subjects), pooled monkey (n=5 monkey subjects), and several DCNN_{IC} models. Each B.O1 signature is shown as a 24-dimensional vector using a color scale; each colored bin corresponds to the system's discriminability of one object against all others that were tested. The color scales span each signature's full performance range, and warm colors indicate lower discriminability. (B) Direct comparison of the B.O1 signatures of a pixel visual system model (top panel) and a DCNN_{IC} visual system model (Inception-v3, bottom panel) against that of the human B.O1 signature. (C) Human-consistency of B.O1 signatures, for each of the tested model visual systems. The black and gray dots correspond to a held-out pool of five human subjects and a pool of five macaque monkey subjects respectively. The shaded area corresponds to the primate zone, a range of consistencies delimited by the estimated human-consistency of a pool of infinitely many monkeys.

between each pair of B.O1 signatures (termed human-consistency, following [Johnson et al., 2002]; the noise adjustment means that a visual system that is identical to the human pool will have an expected human-consistency score of 1.0, even if it has irreducible trial-by-trial stochasticity; see Methods). Figure 3-2C shows the B.O1 human-consistency for each of the tested model visual systems. We additionally tested the behavior of a held-out pool of five human subjects (black dot) and a pool of five macaque monkey subjects (gray dot), and we observed that both yielded B.O1 signatures that were highly human-consistent (human-consistency $\tilde{\rho} = 0.90, 0.97$ for monkey pool and held-out human pool, respectively). We defined a range of human-consistency values, termed the "primate zone" (shaded gray area), delimited by ex-

trapolated human-consistency estimates of large pools of macaques (see Methods, Figure 3-6). We found that the baseline pixel visual system model and the low-level V1 visual system model were not within this zone ($\tilde{\rho} = 0.40, 0.67$ for pixels and V1 models, respectively), while all tested DCNN$_{IC}$ visual system models were either within or very close to this zone. Indeed, we could not reject the hypothesis that DCNN$_{IC}$ models are primate-like ($p = 0.54$, exact test, see Methods).

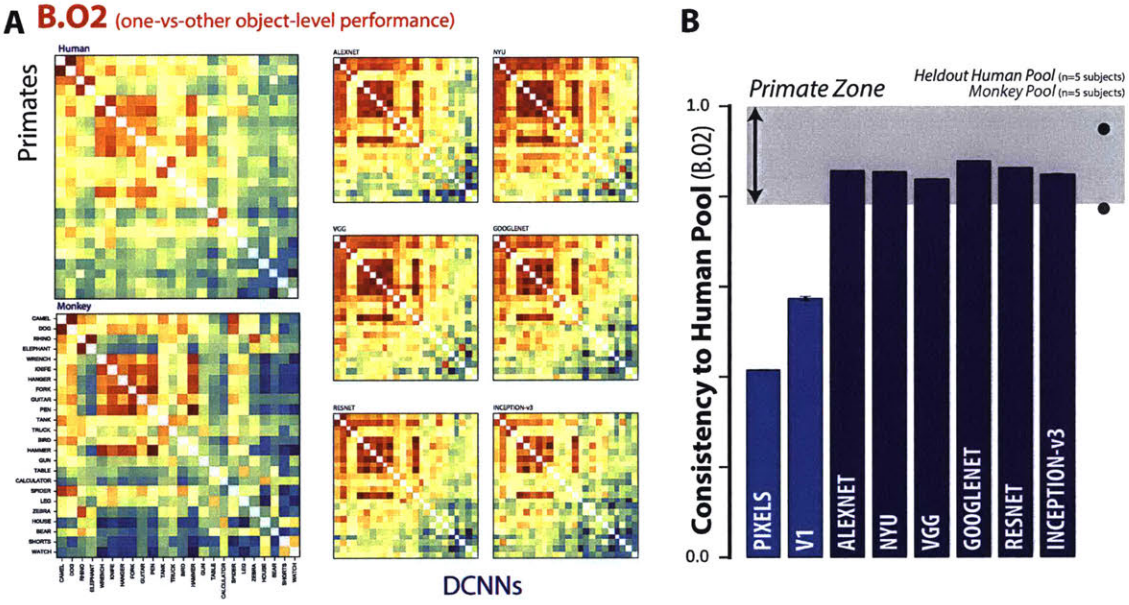

Figure 3-3: One-versus-other object-level (B.O2) signatures for pooled human, pooled monkey, and several DCNN$_{IC}$ models. Each B.O2 signature is shown as a 24x24 symmetric matrices using a color scale, where each bin (i,j) corresponds to the system's discriminability of objects i and j. Color scales similar to (A). (E) Human-consistency of B.O2 signatures for each of the tested model visual systems. Format is identical to (C).

Next, we compared the behavior of the visual systems at a slightly higher level of resolution. Specifically, instead of pooling over all discrimination tasks for each object, we computed the mean discriminability of each of the 276 pairwise discrimination tasks (still pooling over images within each of those tasks). This yielded a symmetric matrix that is referred to here as the B.O2 signature. Figure 3-3A shows the B.O2 signatures of the pooled human, pooled monkey, and several DCNN$_{IC}$ visual system

models as 24x24 symmetric matrices. Each bin (i,j) corresponds to the system's discriminability of objects i and j, where warmer colors indicate lower performance; color scales are not shown but span each signature's full range. We observed strong qualitative similarities between the pairwise object confusion patterns of all of the high level visual systems (e.g. camel and dog are often confused with each other by all three systems). This similarity is quantified in Figure 3-3B, which shows the human-consistency of all examined visual system models with respect to this metric. Similar to the B.O1 metric, we observed that both a pool of macaque monkeys and a held-out pool of humans are highly human-consistent with respect to this metric ($\tilde{\rho} = 0.77, 0.94$ for monkeys, humans respectively). Also similar to the B.O1 metric, we found that all DCNN_{IC} visual system models are highly human-consistent ($\tilde{\rho} > 0.8$) while the baseline pixel visual system model and the low-level V1 visual system model were not ($\tilde{\rho} = 0.41, 0.57$ for pixels, V1 models respectively). Indeed, all DCNN_{IC} visual system models are within the defined "primate zone" of human-consistency, and we could not falsify the hypothesis that DCNN_{IC} models are primate-like ($p = 0.99$, exact test).

Taken together, humans, monkeys, and current DCNN_{IC} models all share similar patterns of object-level behavioral performances (B.O1 and B.O2 signatures) that are not shared with lower-level visual representations (pixels and V1). However, object-level performance patterns do not capture the fact that some images of an object are more challenging than other images of the same object because of interactions of the variation in the object's pose and position with the object's class. To overcome this limitation, we next examined the patterns of behavior at the resolution of individual images on a subsampled set of images where we specifically obtained a large number of behavioral trials to accurately estimate behavioral performance on each image. Note that, from the point of view of the subjects, the behavioral tasks are identical to those already described. We simply aimed to measure and compare their patterns of performance at much higher resolution.

3.2.2 Image-level behavioral comparison

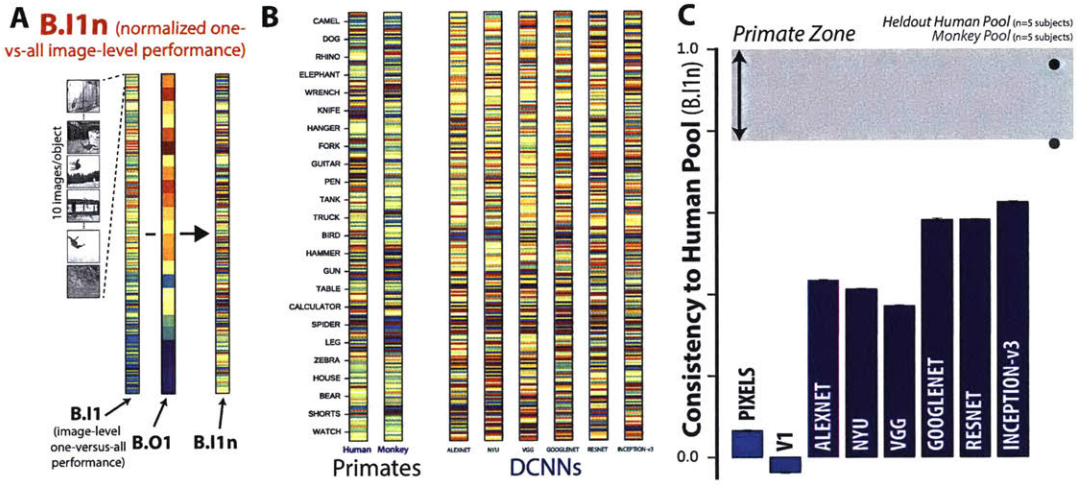

Figure 3-4: (A) Schematic for computing B.I1n. First, the one-versus-all image-level signature (B.I1) is shown as a 240-dimensional vector (24 objects, 10 images/object) using a color scale, where each colored bin corresponds to the system's discriminability of one image against all distractor objects. From this pattern, the normalized one-versus-all image-level signature (B.I1n) is estimated by subtracting the mean performance (d') value over all images of the same object. This normalization procedure isolates behavioral variance that is specifically image-driven but not simply predicted by the object. (B) Normalized one-versus-all object-level (B.I1n) signatures for the pooled human, pooled monkey, and several DCNN_{IC} models. Each B.I1n signature is shown as a 240-dimensional vector using a color scale, formatted as in (A). (C) Human-consistency of B.I1n signatures for each of the tested model visual systems.

To isolate purely image-level behavioral variance, i.e. variance that is not predicted by the object and thus already captured by the B.O1 signature, we computed the normalized image-level signature. This normalization procedure is schematically illustrated in Figure 3-4A which shows that the one-versus-all image-level signature (240-dimensional, 10 images/object) is used to obtain the normalized one-versus-all image-level signature (termed B.I1n, see Behavioral metrics and signatures). Figure 3-4B shows the B.I1n signatures for the pooled human, pooled monkey, and several DCNN_{IC} models as 240 dimensional vectors. Each bin's color corresponds to the discriminability of a single image against all distractor options (after subtraction of object-level discriminability, see Figure 3-4A), where warmer colors indicate lower values; color scales are not shown but span each signature's full range. Fig-

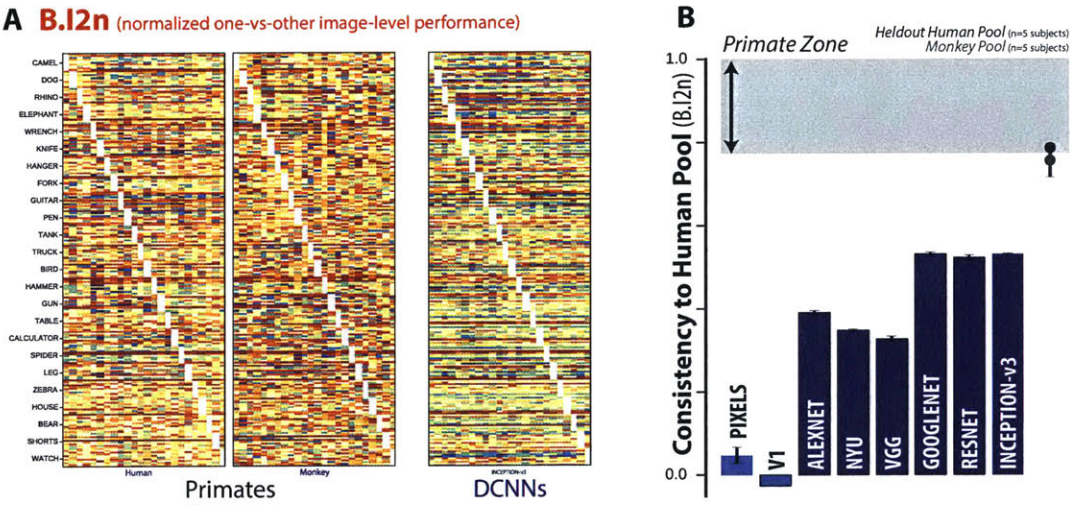

Figure 3-5: (D) Normalized one-versus-other image-level (B.I2n) signatures for pooled human, pooled monkey, and several DCNN$_{IC}$ models. Each B.I2n signature is shown as a 240x24 matrix using a color scale, where each bin (i,j) corresponds to the system's discriminability of image i against distractor object j. (E) Human-consistency of B.I2n signatures for each of the tested model visual systems.

ure 3-4C shows the human-consistency with respect to the B.I1n signature for all tested models. Unlike with object-level behavioral metrics, we now observe a divergence between DCNN$_{IC}$ models and primates. Both the monkey pool and the held-out human pool remain highly human-consistent ($\tilde{\rho} = 0.77, 0.96$ for monkeys, humans respectively), but all DCNN$_{IC}$ models were significantly less human-consistent (Inception-v3: $\tilde{\rho} = 0.62$) and well outside of the defined "primate zone" of B.I1n human-consistency. Indeed, the hypothesis that the human-consistency of DCNN$_{IC}$ models is within the primate zone is strongly rejected ($p = 6.16e-8$, exact test, see Methods).

We can zoom in further by examining not only the overall performance for a given image but also the object confusions for each image, i.e. the additional behavioral variation that is due not only to the test image but to the interaction of that test image with the alternative (incorrect) object choice that is provided after the test image (see Fig. 1B). This is the highest level of behavioral accuracy resolution that our task design allows. In raw form, it corresponds to one-versus-other

image-level confusion matrix, where the size of that matrix is the total number of images by the total number of objects (here, 240x24). Each bin (i,j) corresponds to the behavioral discriminability of a single image i against distractor object j. Again, we isolate variance that is not predicted by object-level performance by subtracting the average performance on this binary task (mean over all images) to convert the raw matrix B.I2 above into the normalized matrix, referred to as B.I2n. Figure 3-5A shows the B.I2n signatures as 240x24 matrices for the pooled human, pooled monkey and top DCNN$_{IC}$ visual system models. Color scales are not shown but span each signature's full range; warmer colors correspond to images with lower performance in a given binary task, relative to all images of that object in the same task. Figure 3-5B shows the human-consistency with respect to the B.I2n metric for all tested visual system models. Extending our observations using B.I1n, we observe a similar divergence between primates and DCNN$_{IC}$ visual system models on the matrix pattern of image-by-distractor difficulties (B.I2n). Specifically, both the monkey pool and held-out human pool remain highly human-consistent ($\tilde{\rho} = 0.75, 0.77$ for monkeys, humans respectively), while all tested DCNN$_{IC}$ models are significantly less human-consistent (Inception-v3: $\tilde{\rho} = 0.53$) falling well outside of the defined "primate zone" of B.I2n human-consistency values. Once again, the hypothesis that the human-consistency of DCNN$_{IC}$ models is within the primate zone is strongly rejected ($p = 3.17e-18$, exact test, see Methods).

3.2.3 Natural subject-to-subject variation

For each behavioral metric (B.O1, BO2, B.I1n, BI2n), we defined a "primate zone" as the range of consistency values delimited by human-consistency estimates ρ_{M_∞} and ρ_{H_∞} as lower and upper bounds respectively. ρ_{M_∞} corresponds to the extrapolated estimate of the human-consistency of a large (i.e. infinitely many subjects) pool of rhesus macaque monkeys. Thus, the fact that a particular tested visual system model falls outside of the primate zone can be interpreted as a failure of that visual system model to accurately predict the behavior of the archetypal human at least as well as

the archetypal monkey.

However, from the above analyses, it is not yet clear whether a visual system model that fails to predict the archetypal human might nonetheless accurately correspond to one or more individual human subjects found within the natural variation of the human population. Given the difficulty of measuring individual subject behavior at the resolution of single images for large numbers of human and monkey subjects, we could not yet directly test this hypothesis. Instead, we examined it indirectly by asking whether an archetypal model—that is a pool that includes an increasing number of model "subjects"—would approach the human pool. We simulated model inter-subject variability by retraining a fixed DCNN architecture with a fixed training image set with random variation in the initial conditions and order of training images. This procedure results in models that can still perform the task but with slightly different learned weight values. We note that this procedure is only one possible choice of generating inter-subject variability within each visual system model type, a choice that is an important open research direction that we do not address here. From this procedure, we constructed multiple trained model instances ("subjects") for a fixed DCNN architecture, and asked whether an increasingly large pool of model "subjects" better captures the behavior of the human pool, at least as well as a monkey pool. This post-hoc analysis was conducted for the most human-consistent DCNN architecture (Inception-v3).

Figure 3-6A shows, for each of the four behavioral metrics, the measured human-consistency of subject pools of varying size (number of subjects n) of rhesus macaque monkeys (black) and ImageNet-trained Inception-v3 models (blue). The human-consistency increases with growing number of subjects for both visual systems across all behavioral metrics. To estimate the expected human-consistency for a pool of infinitely many monkey or model subjects, we fit an exponential function mapping n to the mean human-consistency values and obtained a parameter estimate for the asymptotic value (see Methods). We note that estimated asymptotic values are not

significantly beyond the range of the measured data—the human-consistency of a pool of five monkey subjects reaches within 97% of the human-consistency of an estimated infinite pool of monkeys for all metrics—giving credence to the extrapolated human-consistency values. This analysis suggests that under this model of inter-subject variability, a pool of Inception-v3 subjects accurately capture archetypal human behavior at the resolution of objects (B.O1, B.O2) by our primate zone criterion (see Figure 3-5A, first two panels). In contrast, even a large pool of Inception-v3 subjects still fails at its final asymptote to accurately capture human behavior at the image-level (B.I1n, B.I2n) (Figure 3-5A, last two panels).

3.2.4 Modification of visual system models to try to rescue their human-consistency

Next, we wondered if some relatively simple changes to the DCNN_{IC} visual system models tested here could bring them into better correspondence with the primate visual system behavior (with respect to B.I1n and B.I2n metrics). Specifically, we considered and tested the following modifications to the most human-consistent DCNN_{IC} model visual system (Inception-v3): we (1) changed the input to the model to be more primate-like in its retinal sampling (**Inception-v3+retina**), (2) changed the transformation (aka "decoder") from the internal model feature representation into the behavioral output by augmenting the number of decoder training images or changing the decoder type (**Inception-v3+SVM**, **Inception-v3+classifier-train**), and (3) modified all of the internal filter weights of the model (aka "fine tuning") by augmenting its ImageNet training with additional images drawn from the same distribution as our test images (**Inception-v3+synthetic-fine-tune**). While some of these modifications (e.g. fine-tuning on synthetic images and increasing the number of classifier training images) had the expected effect of increasing mean overall performance (not shown, see Methods), we found that none of these modifications led to a significant improvement in its human-consistency on the behavioral metrics (Figure 3-5B). Thus,

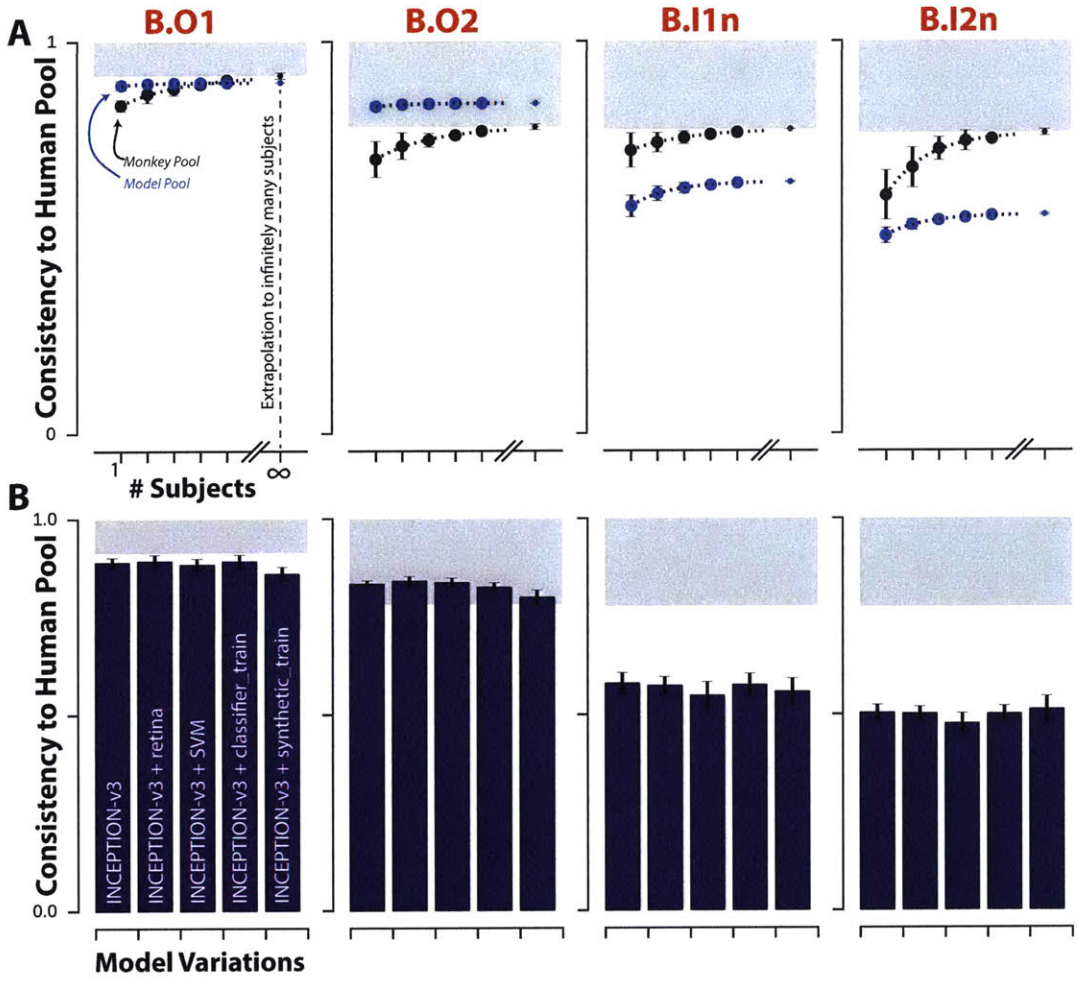

Figure 3-6: Effect of subject pool size and DCNN model modifications on consistency with human behavior. (A) Accounting for natural subject-to-subject variability. For each of the four behavioral metrics, the human-consistency distributions of monkey (blue markers) and model (black markers) pools are shown as a function of the number of subjects in the pool (mean ± SD, over subjects). The human consistency increases with growing number of subjects for all visual systems across all behavioral metrics. The dashed lines correspond to fitted exponential functions, and the parameter estimate (mean ± SE) of the asymptotic value, corresponding to the estimated human-consistency of a pool of infinitely many subjects, is shown at the right most point on each abscissa. (B) Model modifications that aim to rescue the DCNN_{IC} models. We tested several simple modifications (see Methods) to the most human-consistent DCNN_{IC} visual system model (Inception-v3). Each panel shows the resulting human-consistency per modified model (mean ± SD over different model instances, varying in random filter initializations) for each of the four behavioral metrics.

the failure of current DCNN_{IC} models to accurately capture the image-level signatures of primates cannot be rescued by simple modifications on a fixed architecture.

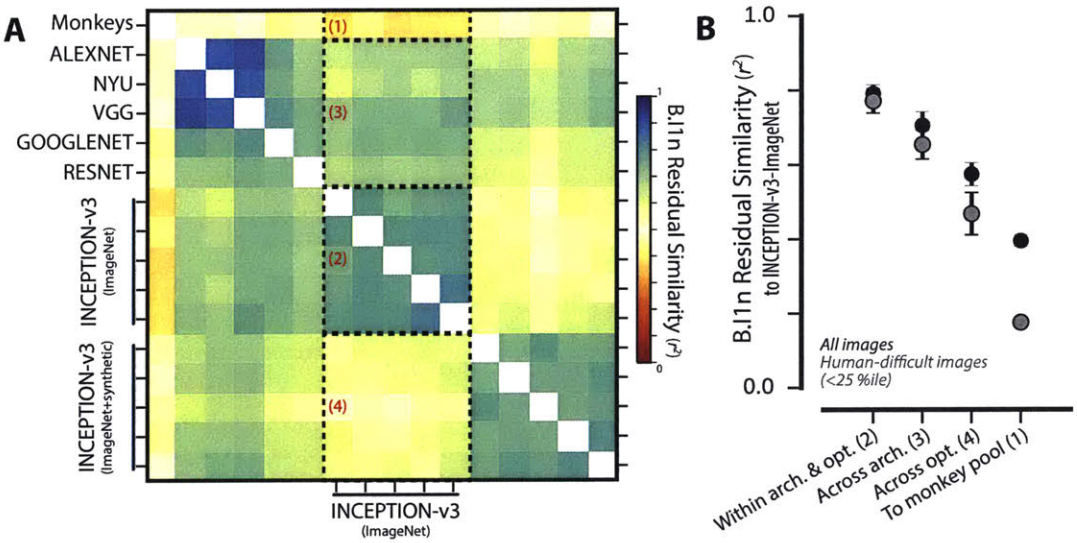

Figure 3-7: Analysis of unexplained human behavioral variance. (A) Residual similarity between all pairs of human visual system models. The color of bin (i,j) indicates the proportion of explainable variance that is shared between the residual signatures of visual systems i and j. For ease of interpretation, we ordered visual system models based on their architecture and optimization procedure and partitioned this matrix into four distinct regions. (B) Summary of residual similarity. For each of the four regions in (a), the similarity to the residuals of Inception-v3 (region 2 in (A)) is shown (mean $\pm$ SD, within each region) for all images (black dots), and for images that humans found to be particularly difficult (gray dots, selected based on held-out human data).

3.2.5 Looking for clues: Image-level comparisons of models and primates

Taken together, the results described above suggest that current DCNN_{IC} visual system models fail to accurately capture the image-level signatures of humans and monkeys. To further examine this failure in the hopes of providing clues for model improvement, we examined the image-level residual signatures of all the visual system models, relative to the pooled human. For each model, we computed its residual

signature as the difference (positive or negative) of a linear least squares regression of the model signature on the corresponding human signature. For this analysis, we focused on the B.I1n metric as it showed a clear divergence of $DCNN_{IC}$ models and primates, and the behavioral residual can be interpreted based only on the test images (whereas B.I2n depends on the interaction between test images and distractor choice).

We first asked to what extent the residual signatures are shared between different visual system models. Figure 3-7A shows the similarity between the residual signatures of all pairs of models; the color of bin (i,j) indicates the proportion of explainable variance that is shared between the residual signatures of visual systems i and j. For ease of interpretation, we ordered visual system models based on their architecture and optimization procedure and partitioned this matrix into four distinct regions. Each region compares the residuals of a "source" model group with fixed architecture and optimization procedure (five Inception-v3 models optimized for categorization on ImageNet, varying only in initial conditions and training image order) to a "target" model group. The target groups of models for each of the four regions are: 1) the pooled monkey, 2) other $DCNN_{IC}$ models from the source group, 3) $DCNN_{IC}$ models that differ in architecture but share the optimization procedure of the source group models and 4) $DCNN_{IC}$ models that differ slightly using an augmented optimization procedure but share the architecture of the source group models. Figure 3-7B shows the mean ($\pm$SD) variance shared in the residuals averaged within these four regions for all images (black dots), as well as for images that humans found to be particularly difficult (gray dots, selected based on held-out human data, see Methods). First, consistent with the results shown in Figure 3-4, we note that the residual signatures of this particular $DCNN_{IC}$ model are not well shared with the pooled monkey ($r^2 = 0.39$ in region 1), and this phenomenon is more pronounced for the images that humans found most difficult ($r^2 = 0.17$ in region 1). However, this relatively low correlation between model and primate residuals is not indicative of spurious model residuals, as the model residual signatures were highly reliable between different instances of this fixed $DCNN_{IC}$ model, across random training ini-

tializations (region 2: $r^2 = 0.79, 0.77$ for all and most difficult images, respectively). Interestingly, residual signatures were still largely shared with other DCNN$_{IC}$ models with vastly different architectures (region 3: $r^2 = 0.70, 0.65$ for all and most difficult images, respectively). However, residual signatures were more strongly altered when the visual training diet of the same architecture was altered (region 4: $r^2 = 0.57, 0.46$ for all and most difficult images respectively, cf. region 3). Taken together, these results indicate that the images where DCNN$_{IC}$ visual system models diverged from humans (and monkeys) were not spurious but were rather highly reliable across different model architectures, demonstrating that current DCNN$_{IC}$ models systematically and similarly diverge from primates.

To look for clues for model improvement, we asked what, if any, characteristics of images might account for this divergence of models and primates. We regressed the residual signatures of DCNN$_{IC}$ models on four different image attributes (corresponding to the size, eccentricity, pose, and contrast of the object in each image). We used multiple linear regressions to predict the model residual signatures from all of these image attributes, and also considered each attribute individually using simple linear regressions. Figure 3-8A shows example images (sampled from the full set of 2400 images) with increasing attribute value for each of these four image attributes. While the DCNN$_{IC}$ models were not directly optimized to display primate-like performance dependence on such attributes, we observed that the Inception-v3 visual system model nonetheless exhibited qualitatively similar performance dependencies as primates (see Figure 3-8B). For example, humans (black), monkeys (gray) and the Inception-v3 model (blue) all performed better, on average, for images in which the object is in the center of gaze (low eccentricity) and large in size. Furthermore, all three systems performed better, on average, for images when the pose of the object was closer to the canonical pose (see Figure 3-8B). The similarity of the patterns in Figure 3-8B between primates and the DCNN$_{IC}$ visual system models is not perfect but is striking, particularly in light of the fact that these models were not optimized to produce these patterns. However, this similarity is analogous to the similarity

in the B.O1 and B.O2 metrics in that it only holds on average over many images. Looking more closely at the image-by-image comparison, we again found that the DCNN$_{IC}$ models failed to capture a large portion of the image-by-image variation. In particular, Figure 3-8C shows the proportion of variance explained by specific image attributes for the residual signatures of monkeys (black) and DCNN$_{IC}$ models (blue). We found that, taken together, all four of these image attributes explained only $\sim$ 10% of the variance in DCNN$_{IC}$ residual signatures, and each individual attribute could explain at most a small amount of residual variance (<5% of the explainable variance). In sum, these analyses show that some behavioral effects that might provide intuitive clues to modify the DCNN$_{IC}$ models are already in place in those models (e.g. a dependence on eccentricity). But the quantitative image-by-image analyses of the remaining unexplained variance (Figure 3-8C) argue that the DCNN$_{IC}$ visual system models' failure to capture primate image-level signatures cannot be further accounted for by these simple image attributes and likely stem from other factors.

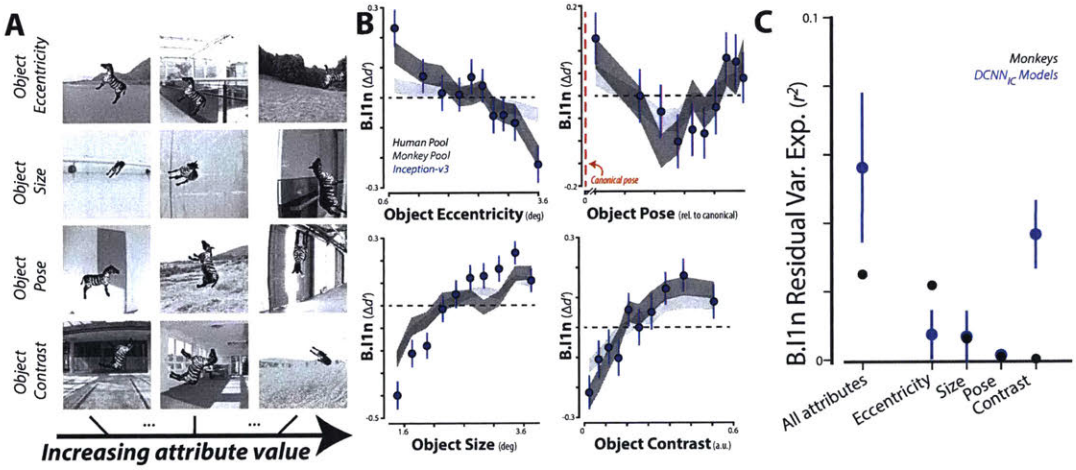

Figure 3-8: Dependence of primate and DCNN$_{IC}$ model behavior on image attributes. (A) Example images with increasing attribute value, for each of the four pre-defined image attributes (see Methods). (B) Dependence of performance (B.I1n) as a function of four image attributes, for humans, monkeys and a DCNN$_{IC}$ model (Inception-v3). (C) Proportion of explainable variance of the residual signatures of monkeys (black) and DCNN$_{IC}$ models (blue) that is accounted for by each of the pre-defined image attributes. Error-bars correspond to SD over trial re-sampling for monkeys, and over different models for DCNN$_{IC}$ models.

3.3 Discussion

The current work was motivated by the broad scientific goal of discovering models that quantitatively explain the neuronal mechanisms underlying primate invariant object recognition behavior. To this end, previous work had shown that specific artificial neural network models (ANNs), drawn from a large family of deep convolutional neural networks (DCNNs) and optimized to achieve high levels of object categorization performance on large-scale image-sets, capture a large fraction of the variance in primate visual recognition behaviors [Rajalingham et al., 2015, Jozwik et al., 2016, Kheradpisheh et al., 2016, Kubilius et al., 2016, Peterson et al., 2016, Wallis et al., 2017], and the internal hidden neurons of those same models also predict a large fraction of the image-driven response variance of brain activity at multiple stages of the primate ventral visual stream [Seibert et al., 2016, Cadieu et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014, Guclu and van Gerven, 2015, Cichy et al., 2016, Hong et al., 2016, Seibert et al., 2016, Cadena et al., 2017, Wen et al., 2017]. For clarity, we here referred to this sub-family of models as DCNN_{IC} (to denote ImageNet-Categorization training), so as to distinguish them from all possible models in the DCNN family, and more broadly, from the super-family of all ANNs. In this work, we directly compared leading DCNN_{IC} models to primates (humans and monkeys) with respect to their behavioral signatures at both object and image level resolution in the domain of core object recognition. In order to do so, we measured and characterized primate behavior at larger scale and higher resolution than previously possible. We first replicate prior work (Rajalingham et al., 2015) showing that, at the object level, DCNN_{IC} models produce statistically indistinguishable behavior from primates, and we extend that work by showing that these models also match the average primate sensitivities to object contrast, eccentricity, size, and pose, a noteworthy similarity in light of the fact that these models were not optimized to produce these performance patterns. However, our primary novel result is that, examining behavior at the higher resolution of individual images, all leading DCNN_{IC} models failed to replicate the image-level behavioral signatures of primates. An important

related claim is that rhesus monkeys are more consistent with the archetypal human than any of the tested DCNN_{IC} models (at the image-level).

While it had previously been shown that DCNN_{IC} models can diverge from human behavior on specifically chosen adversarial images [Szegedy et al., 2013], a strength of our work is that we did not optimize images to induce failure but instead randomly sampled the image generative parameter space broadly. As such, our results highlight a general, rather than adversarial-induced, failure of DCNN_{IC} models to fully capture the neural mechanisms underlying primate core object recognition behavior. Furthermore, we showed that this failure of current DCNN_{IC} models cannot be explained by simple image attributes and cannot be rescued by simple model modifications (input image sampling, model training, and classifier variations). Taken together, these results suggest that new ANN models are needed to more precisely capture the neural mechanisms underlying primate object vision.

With regards to new ANN models, we can attempt to make prospective inferences about future possible DCNN_{IC} models from the data presented here. Based on the observed distribution of image-level human-consistency values for the DCNN_{IC} models tested here, we infer that yet untested model instances sampled identically (i.e. from the DCNN_{IC} model sub-family) are very likely to have similarly inadequate image-level human-consistency. While we cannot rule out the possibility that at least one model instance within the DCNN_{IC} sub-family would fully match the image-level behavioral signatures, the probability of sampling such a model is vanishingly small ($p < 1e-17$ for B.I2n human-consistency, estimated using exact test using Gaussian kernel density estimation, see Methods, Results). An important caveat of this inference is that we may have a biased estimate of the human-consistency distribution of this model sub-family, as we did not exhaustively sample the sub-family. In particular, if the model sampling process is non-stationary over time (e.g. increases in computational power over time allows larger models to be successfully trained), the human-consistency of new (i.e. yet to be sampled) models may lie outside the

currently estimated distribution. Consistent with the latter, we observed that current DCNN_{IC} cluster into two distinct "generations" separated in time (before/after the year 2015; e.g. Inception-v3 improves over AlexNet though both lie outside the primate zone in Figure 3-4). Thus, following this trend, it is possible that the evolution of "next-generation" models within the DCNN_{IC} sub-family could meet our criteria for successful matching primate-like behavior.

Alternatively, it is possible—and we think likely—that future DCNN_{IC} models will also fail to capture primate-like image-level behavior, suggesting that either the architectural limitations (e.g. convolutional, feed-forward) and/or the optimization procedure (including the diet of visual images) that define this model sub-family are fundamentally limiting. Thus, ANN model sub-families utilizing different architectures (e.g. recurrent neural networks) and/or optimized for different behavioral goals (e.g. loss functions other than object classification performance, and/or images other than category-labeled ImageNet images) may be necessary to accurately capture primate behavior. To this end, we propose that testing even individual changes to the DCNN_{IC} models—each creating a new ANN model sub-family—may be the best way forward, because DCNN_{IC} models currently offer the best explanations (in a predictive sense) of both the behavioral and neural phenomena of core object recognition.

To reach that goal of finding a new ANN model sub-family that is a better mechanistic model of the primate ventral visual stream, we propose that even larger-scale, high-resolution behavioral measurements, such as expanded versions of the patterns of image-level performance presented here, could serve as a useful top-down optimization guide. Not only do these high-resolution behavioral signatures have the statistical power to reject the currently leading ANN models, but they can also be efficiently collected at very large scale, in contrast to other guide data (e.g. large-scale neuronal measurements). Indeed, current technological tools for high-throughput psychophysics in humans and monkeys (e.g. Amazon Mechanical Turk for humans, Monkey Turk for rhesus monkeys) enable time- and cost-efficient collection of large-

scale behavioral datasets, such as the $\sim$ 1 million behavioral trials obtained for the current work. These systems trade off an increase in efficiency with a decrease in experimental control. For example, we did not impose experimental constraints on subjects' acuity and we can only infer likely head and gaze position. Previous work has shown that patterns of behavioral performance on object recognition tasks from in-lab and online subjects were equally reliable and virtually identical [Majaj et al., 2015], but it is not yet clear to what extent this holds at the resolution of individual images, as one might expect that variance in performance across images is more sensitive to precise head and gaze location. For this reason, we here refrain from making strong inferences from small behavioral differences, such as the small difference between humans and monkeys. Nevertheless, we argue that this sacrifice in exact experimental control while retaining sufficient power for model comparison is a good tradeoff for efficiently collecting large behavioral datasets toward the goal of constraining future models of the primate ventral visual stream.

3.4 Methods

3.4.1 Visual images

We examined basic-level, core object recognition behavior using a set of 24 broadly-sampled objects that we previously found to be reliably labeled by independent human subjects, based on the definition of basic-level proposed by [Rosch et al., 1976]. These images are identical to those used in Chapter 2; Figure 2-1 shows the full list of 24 objects, with two example images of each object.

Because all of the images were generated from synthetic 3D object models, we had explicit knowledge of the viewpoint parameters (position, size, and pose) for each object in each image, as well as perfect segmentation masks. Taking advantage of this feature, we characterized each image based on these high-level attributes, consisting

of size, eccentricity, relative pose and contrast of the object in the image. The size and eccentricity of the object in each image were computed directly from the corresponding viewpoint parameters, under the assumption that the entire image would subtend 6ą at the center of visual gaze (±3° in both azimuth and elevation; see below). For each synthetic object, we first defined its "canonical" 3D pose vector, based on independent human judgments. To compute the relative pose attribute of each image, we estimated the difference between the object's 3D pose and its canonical 3D pose. Pose differences were computed as distances in unit quaternion representations: the 3D pose (r_{xy}, r_{xz}, r_{yz}) was first converted into unit quaternions, and distances between quaternions q_1, q_2 were estimated as $cos^{-1}|q_1 \cdot q_2|$ [Huynh, 2009]. To compute the object contrast, we measured the absolute difference between the mean of the pixel intensities corresponding to the object and the mean of the background pixel intensities in the vicinity of the object (specifically, within 25 pixels of any object pixel, analogous to computing the local foreground-background luminance difference of a foreground object in an image). Figure 3-8A shows example images with varying values for the four image attributes.

3.4.2 Core object recognition behavioral paradigm

Core object discrimination is defined as the ability to discriminate between two or more objects in visual images presented under high view uncertainty in the central visual field ($\sim 10°$), for durations that approximate the typical primate, free-viewing fixation duration (200 ms) [DiCarlo and Cox, 2007, DiCarlo et al., 2012]. As in our previous work [Rajalingham et al., 2015], the behavioral task paradigm consisted of a interleaved set of binary discrimination tasks. Each binary discrimination task is an object discrimination task between a pair of objects (e.g. elephant vs. bear). Each such binary task is balanced in that the test image is equally likely (50%) to be of either of the two objects. On each trial, a test image is presented, followed by a choice screen showing canonical views of the two possible objects (the object that was not displayed in the test image is referred to as the "distractor" object, but

note that objects are equally likely to be distractors and targets). Here, 24 objects were tested, which resulted in 276 binary object discrimination tasks. To neutralize feature attention, these 276 tasks are randomly interleaved (trial by trial), and the global task is referred to as a basic-level, core object recognition task paradigm.

Testing human behavior

All human behavioral data presented here were collected from 1476 human subjects on Amazon Mechanical Turk (MTurk) performing the task paradigm described above. Subjects were instructed to report the identity of the foreground object in each presented image from among the two objects presented on the choice screen (Fig 1B). Because all 276 tasks were interleaved randomly (trial-by-trial), subjects could not deploy feature attentional strategies specific to each object or specific to each binary task to process each test image.

Figure 3-2A illustrates the time course of each behavioral trial, for a particular object discrimination task (zebra versus dog). Each trial initiated with a central black point for 500 ms, followed by 100 ms presentation of a test image containing one foreground object presented under high variation in viewing parameters and overlaid on a random background, as described above (see Visual images above). Immediately after extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were shown to the left and right. One of these two objects was always the same as the object that generated the test image (i.e., the correct object choice), and the location of the correct object (left or right) was randomly chosen on each trial. After clicking on one of the choice images, the subject was queued with another fixation point before the next test image appeared. No feedback was given; human subjects were never explicitly trained on the tasks. Under assumptions of typical computer ergonomics, we estimate that images were presented at $6-8^{\circ}$ of visual angle at the center of gaze, and the choice object images were presented at $\pm 6-8^{\circ}$ of eccentricity along the horizontal meridian.

We measured human behavior using the online Amazon MTurk platform (see Figure 3-2B), which enables efficient collection of large-scale psychophysical data from crowd-sourced "human intelligence tasks" (HITs). The reliability of the online MTurk platform has been validated by comparing results obtained from online and in-lab psychophysical experiments [Majaj et al., 2015, Rajalingham et al., 2015]. We pooled 927,296 trials from 1472 human subjects to characterize the aggregate human behavior, which we refer to as the "pooled" human (or "archetypal" human). Each human subject performed only a small number of trials ($\sim$ 150) on a subset of the images and binary tasks. All 2400 images were used for behavioral testing, but in some of the HITs, we biased the image selection towards the 240 primary test images (1424$\pm$70 trials/image on this subsampled set, versus 271$\pm$93 trials/image on the remaining images, mean $\pm$ SD) to efficiently characterize behavior at image level resolution. Images were randomly drawn such that each human subject was exposed to each image a relatively small number of times (1.5$\pm$2.0 trials/image per subject, mean $\pm$ SD), in order to mitigate potential alternative behavioral strategies (e.g. "memorization" of images) that could arise from a finite image set. Behavioral signatures at the object-level (B.O1, B.O2, see Behavioral metrics and signatures) were measured using all 2400 test images, while image-level behavioral signatures (B.I1n, B.I2n, see Behavioral metrics and signatures) were measured using the 240 primary test images. (We observed qualitatively similar results using those metrics on the full 2400 test images, but we here focus on the primary test images as the larger number of trials leads to lower noise levels).

Five other human subjects were separately recruited on MTurk to each perform a large number of trials on the same images and tasks (53,097$\pm$15,278 trials/subject, mean $\pm$ SD). Behavioral data from these five subjects was not included in the characterization of the pooled human described above, but instead aggregated together to characterize a distinct held-out human pool. For the scope of the current work, this held-out human pool—which largely replicated all behavioral signatures of the larger

archetypal human—served as an independent validation of our human behavioral measurements.

Testing monkey behavior

Five adult male rhesus macaque monkeys (Macaca mulatta, subjects M, Z, N, P, B) were tested on the same basic-level, core object recognition task paradigm described above, with minor modification as described below. All procedures were performed in compliance with National Institutes of Health guidelines and the standards of the Massachusetts Institute of Technology Committee on Animal Care and the American Physiological Society. To efficiently characterize monkey behavior, we used a novel home-cage behavioral system developed in our lab (termed MonkeyTurk, see Fig. 1C). This system leveraged a tablet touchscreen (9" Google Nexus or 10.5" Samsung Galaxy Tab S) and used a web application to wirelessly load the task and collect the data (code available at https://github.com/dicarlolab/mkturk). Analogous to the online Amazon Mechanical Turk, which allows for efficient psychophysical assays of a large number (hundreds) of human users in their native environments, MonkeyTurk allowed us to test many monkey subjects simultaneously in their home environment. Each monkey voluntarily initiated trials, and each readily performed the task a few hours each day that the task apparatus was made available to it. At an average rate of $\sim$ 2000 trials per day per monkey, we collected a total of 836,117 trials from the five monkey subjects over a period of $\sim$ 3 months.

Monkey training is described in detail elsewhere [Rajalingham et al., 2015]. Briefly, all monkeys were initially trained on the match-test-image-to-object rule using other images and were also trained on discriminating the particular set of 24 objects tested here using a separate set of training images rendered from these objects, in the same manner as the main testing images. Two of the monkeys subjects (Z and M) were previously trained in the lab setting, and the remaining three subjects were trained using MonkeyTurk directly in their home cages and did not have significant prior lab exposure. Once monkeys reached saturation performance on training images, we

began the behavioral testing phase to collect behavior on test images. Monkeys did improve throughout the testing phase, exhibiting an increase in performance between the first and second half of trials of 4%±0.9% (mean ± SEM over five monkey subjects). However, the image-level behavioral signatures obtained from the first and the second halves of trials were highly correlated to each other (B.I1 noise-adjusted correlation of 0.85±0.06, mean ± SEM over five monkey subjects, see Behavioral metrics and signatures below), suggesting that monkeys did not significantly alter strategies (e.g. did not "memorize" images) throughout the behavioral testing phase.

The monkey task paradigm was nearly identical to the human paradigm (see Figure 3-2B), with the exception that trials were initiated by touching a white "fixation" circle horizontally centered on the bottom third of the screen (to avoid occluding centrally-presented test images with the hand). This triggered a 100ms central presentation of a test image, followed immediately by the presentation of the two choice images (Fig. 1B, location of correct choice randomly assigned on each trial, identical to the human task). Unlike the main human task, monkeys responded by directly touching the screen at the location of one of the two choice images. Touching the choice image corresponding to the object shown in the test image resulted in the delivery of a drop of juice through a tube positioned at mouth height (but not obstructing view), while touching the distractor choice image resulted in a three second timeout. Because gaze direction typically follows the hand during reaching movements, we assumed that the monkeys were looking at the screen during touch interactions with the fixation or choice targets. In both the lab and in the home cage, we maintained total test image size at $\sim 6^\circ$ of visual angle at the center of gaze, and we took advantage of the retina-like display qualities of the tablet by presenting images pixel matched to the display (256 x 256 pixel image displayed using 256 x 256 pixels on the tablet at a distance of 8 inches) to avoid filtering or aliasing effects.

As with Mechanical Turk testing in humans, MonkeyTurk head-free home-cage testing enables efficient collection of reliable, large-scale psychophysical data but it

likely does not yet achieve the level of experimental control that is possible in the head-fixed laboratory setting. However, we note that when subjects were engaged in home-cage testing, they reliably had their mouth on the juice tube and their arm positioned through an armhole. These spatial constraints led to a high level of head position trial-by-trial reproducibility during performance of the task paradigm. Furthermore, when subjects were in this position, they could not see other animals as the behavior box was opaque, and subjects performed the task at a rapid pace 40 trials/minute suggesting that they were not frequently distracted or interrupted. The location of the upcoming test image (but not the location of the object within that test image) was perfectly predictable at the start of each behavioral trial, which likely resulted in a reliable, reproduced gaze direction at the moment that each test image was presented. The relatively short—but natural and high performing [Cadieu et al., 2014]—test image duration (100 ms) ensured that saccadic eye movements were unlike to influence test image performance (as they generally take $\sim$ 200 ms to initiate in response to the test image, and thus well after the test image has been extinguished).

Testing model behavior

We tested a number of different deep convolutional neural network (DCNN) models on the exact same images and tasks as those presented to humans and monkeys. Importantly, our core object recognition task paradigm is closely analogous to the large-scale ImageNet 1000-way object categorization task for which these networks were optimized and thus expected to perform well. We focused on publicly available DCNN model architectures that have proven highly successful with respect to this computer vision benchmark over the past five years: AlexNet [Krizhevsky et al., 2012], NYU [Zeiler and Fergus, 2014], VGG [Simonyan and Zisserman, 2014], GoogleNet [Szegedy et al., 2013]), Resnet [He et al., 2016], and Inception-v3 [Szegedy et al., 2013]. As this is only a subset of possible DCNN models, we refer to these as the DCNN$_{IC}$ (to denote ImageNet-Categorization) visual system model sub-family. For each of the publicly available model architectures, we first used ImageNet-categorization-trained

model instances, either using publicly available trained model instances or training them to saturation on the 1000-way classification task in-house. Training took several days on 1-2 GPUs.

We then performed psychophysical experiments on each ImageNet-trained DCNN model to characterize their behavior on the exact same images and tasks as humans and monkeys. We first adapted these ImageNet-trained models to our 24-way object recognition task by re-training the final class probability layer (initially corresponding to the probability output of the 1000-way ImageNet classification task) while holding all other layers fixed. In practice, this was done by extracting features from the penultimate layer of each DCNN_{IC} (i.e. top-most prior to class probability layer), on the same images that were presented to humans and monkeys, and training back-end multi-class logistic regression classifiers to determine the cross-validated output class probability for each image. This procedure is illustrated in Figure 3-2C. To estimate the hit rate of a given image in a given binary classification task, we renormalized the 24-way class probabilities of that image, considering only the two relevant classes, to sum to one. Object-level and image-level behavioral metrics were computed based on these hit rate estimates (as described in Behavioral metrics and signatures below). Importantly, this procedure assumes that the model "retina" layer processes the central 6° of the visual field. It also assumes that linear discriminants ("readouts") of the model's top feature layer are its behavioral output (as intended by the model designers). Manipulating either of these choices (e.g. resizing the input images such that they span only part of the input layer, or building linear discriminates for behavior using a different model feature layer) would result in completely new, testable ANN models that we do not test here.

From these analyses, we selected the most human-consistent DCNN_{IC} architecture (Inception-v3, see Behavioral consistency below), fixed that architecture, and then performed post-hoc analyses in which we varied: the input image sampling, the initial parameter settings prior to training, the filter training images, the type of

classifiers used to generate the behavior from the model features, and the classifier training images. To examine input image sampling, we re-trained the Inception-v3 architecture on images from ImageNet that were first spatially filtered to match the spatial sampling of the primate retina (i.e. an approximately exponential decrease in cone density away from the fovea) by effectively simulating a fish-eye transformation on each image. These images were at highest resolution at the "fovea" (i.e. center of the image) with gradual decrease in resolution with increasing eccentricity. To examine the analog of "inter-subject variability", we constructed multiple trained model instances ("subjects"), where the architecture and training images were held fixed (Inception-v3 and ImageNet, respectively) but the model filter weights initial condition and order of training images were randomly varied for each model instance. Importantly, this procedure is only one possible choice for simulating inter-subject variability for DCNN models, a choice that is an important open research direction that we do not address here. To examine the effect of model training, we fine-tuned an ImageNet-trained Inception-v3 model on a synthetic image set consisting of 6.9 million images of 1049 objects (holding out 50,000 images for model validation). These images were generated using the same rendering pipeline as our test images, but the objects were non-overlapping with the 24 test objects presented here. As expected, fine-tuning on synthetic images led to an overall increase in performance of 5%. We tested the effect of different classifiers to generate model behavior by testing both multi-class logistic regression and support vector machine classifiers. Additionally, we tested the effect of varying the number of training images used to train those classifiers (20 versus 50 images per class).

3.4.3 Behavioral metrics and signatures

To characterize the behavior of any visual system, we here introduce four behavioral (B) metrics of increasing richness, requiring increasing amounts of data to measure reliably. Each behavioral metric computes a pattern of unbiased behavioral performance, using a sensitivity index: $d' = Z(\text{HitRate}) - Z(\text{FalseAlarmRate})$, where Z is

the inverse of the cumulative Gaussian distribution. The various metrics differ in the resolution at which hit rates and false alarm rates are computed. Table 1 summarizes the four behavioral metrics, varying the hit-rate resolution (object-level or image-level) and the false-alarm resolution (one-versus-all or one-versus-other). When each metric is applied to the behavioral data of a visual system—biological or artificial—we refer to the result as one behavioral "signature" of that system. Note that we do not consider the signatures obtained here to be the final say on the behavior of these biological or artificial systems—in the terms defined here, new experiments using new objects/images but the same metrics would produce additional behavioral signatures.

Metric	Equation	Hit Rate	False Alarm Rate
B.O1	$O_1(i) = Z(HR_i) - Z(FAR_i)$	% trials when images of object i were correctly labeled as object i	% trials when any image was incorrectly labeled as object i.
B.O2	$O_2(i,j) = Z(HR_{i,j}) - Z(FAR_{i,j})$	% trials when images of object i were correctly labeled as object i when presented against distractor object j	% trials when images of object j were incorrectly labeled as object i
B.I1	$I_1(i) = Z(HR_{ii}) - Z(FAR_{ii})$	% trials when image ii was correctly classified as object i	% trials when any image was incorrectly labeled as object i.
B.I2	$I_2(i) = Z(HR_{ii,j}) - Z(FAR_{i,j})$	% trials when image ii was correctly classified as object i, when presented against distractor object j	% trials when images of object j were incorrectly labeled as object i

Table 3.1: Definition of behavioral performance metrics. The first column provides the name, abbreviation, dimensions, and equations for each of the raw performance metrics. The next two columns provide the definitions for computing the hit rate (HR) and false alarm rate (FAR) respectively.

The four behavioral metrics we chose are as follows: First, the one-versus-all object-level performance metric (termed B.O1) estimates the discriminability of each object from all other objects, pooling across all distractor object choices. Since we here tested 24 objects, the resulting B.O1 signature has 24 independent values. Second, the one-versus-other object-level performance metric (termed B.O2) estimates the discriminability of each specific pair of objects, or the pattern of pairwise object

confusions. Since we here tested 276 interleaved binary object discrimination tasks, the resulting B.O2 signature has 276 independent values (the off-diagonal elements on one half of the 24x24 symmetric matrix). Third, the one-versus-all image-level performance metric (termed B.I1) estimates the discriminability of each image from all other objects, pooling across all possible distractor choices. Since we here focused on the primary image test set of 240 images (10 per object, see above), the resulting B.I1 signature has 240 independent values. Fourth, the one-versus-other image-level performance metric (termed B.I2) estimates the discriminability of each image from each distractor object. Since we here focused on the primary image test set of 240 images (10 per object, see above) with 23 distractors, the resulting B.I2 signature has 5520 independent values.

Naturally, object-level and image-level behavioral signatures are tightly linked. For example, images of a particularly difficult-to-discriminate object would inherit lower performance values on average as compared to images from a less difficult-to-discriminate object. To isolate the behavioral variance that is specifically driven by image variation and not simply predicted by the objects (and thus already captured by B.O1 and B.O2), we defined normalized image-level behavioral metrics (termed B.I1n, B.I2n) by subtracting the mean performance values over all images of the same object and task. This process is schematically illustrated in Figure 3-4A. We note that the resulting normalized image-level behavioral signatures capture a significant proportion of the total image-level behavioral variance in our data (e.g. 52%, 58% of human B.I1 and B.I2 variance is driven by image variation, independent of object identity). In this study, we use these normalized metrics for image-level behavioral comparisons between models and primates (see Results).

3.4.4 Behavioral Consistency

To quantify the similarity between a model visual system and the human visual system with respect to a given behavioral metric, we used a measure called the "human-

consistency" as previously defined [Johnson et al., 2002]. Human-consistency ($\tilde{\rho}$) is computed, for each of the four behavioral metrics, as a noise-adjusted correlation of behavioral signatures [DiCarlo and Johnson, 1999]. For each visual system, we randomly split all behavioral trials into two equal halves and applied each behavioral metric to each half, resulting in two independent estimates of the system's behavioral signature with respect to that metric. We took the Pearson correlation between these two estimates of the behavioral signature as a measure of the reliability of that behavioral signature given the amount of data collected, i.e. the split-half internal reliability. To estimate the human-consistency, we computed the Pearson correlation over all the independent estimates of the behavioral signature from the model (m) and the human (h), and we then divide that raw Pearson correlation by the geometric mean of the split-half internal reliability of the same behavioral signature measured for each system:

$$\tilde{\rho}(h,m) = \frac{\rho_{h,m}}{\sqrt{\rho_{h,h} \times \rho_{m,m}}}$$

Since all correlations in the numerator and denominator were computed using the same amount of trial data (exactly half of the trial data), we did not need to make use of any prediction formulas (e.g. extrapolation to larger number of trials using Spearman-Brown prediction formula, as in Chapter 2). This procedure was repeated 10 times with different random split-halves of trials. Our rationale for using a reliability-adjusted correlation measure for human-consistency was to account for variance in the behavioral signatures that arises from "noise," i.e., variability that is not replicable by the experimental condition (image and task) and thus that no model can be expected to predict [DiCarlo and Johnson, 1999, Johnson et al., 2002]. In sum, if the model (m) is a replica of the archetypal human (h), then its expected human-consistency is 1.0, regardless of the finite amount of data that are collected.

3.4.5 Characterization of Residuals

In addition to measuring the similarity between the behavioral signatures of primates and models (using human-consistency analyses, as described above), we examined the corresponding differences, termed "residual signatures." Each candidate visual system model's residual signature was estimated as the residual of a linear least squares regression of the model's signature on the corresponding human signature (with both slope and intercept as free parameters). This procedure effectively captures the differences between human and model signatures after accounting for overall performance differences. Residual signatures were estimated on disjoint split-halves of trials, repeating 10 times with random trial permutations. Residuals were computed with respect to the normalized one-versus-all image-level performance metric (B.I1n) as this metric showed a clear difference between DCNN_{IC} models and primates, and the behavioral residual can be interpreted based only the test images (i.e. we can assign a residual per image).

To examine the extent to which the difference between each model and the archetypal human is reliably shared across different models, we measured the Pearson correlation between the residual signatures of pairs of models. Residual similarity was quantified as the proportion of shared variance, defined as the square of the noise-adjusted correlation between residual signatures (the noise-adjustment was done as defined in equation above). Correlations of residual signatures were always computed across distinct split-halves of data, to avoid introducing spurious correlations from subtracting common noise in the human data. We measured the residual similarity between all pairs of tested models, holding both architecture and optimization procedure fixed (between instances of the ImageNet-categorization trained Inception-v3 model, varying in filter initial conditions), varying the architecture while holding the optimization procedure fixed (between all tested ImageNet-categorization trained DCNN architectures), and holding the architecture fixed while varying the optimization procedure (between ImageNet-categorization trained Inception-v3 and

synthetic-categorization fine-tuned Inception-v3 models). This analysis addresses not only the reliability of the failure of DCNN_{IC} models to predict human behavior (deviations from humans), but also the relative importance of the characteristics defining similarities within the model sub-family (namely, the architecture and the optimization procedure). We first performed this analysis for residual signatures over the 240 primary test images, and subsequently zoomed in on subsets of images that humans found to be particularly difficult. This image selection was made relative to the distribution of image-level performance of held-out human subjects (B.I1 metric from five subjects); difficult images were defined as ones with performance below the 25th percentile of this distribution.

To examine whether the difference between each model and humans can be explained by simple human-interpretable stimulus attributes, we regressed each DCNN_{IC} model's residual signature on image attributes (object size, eccentricity, pose, and contrast). Briefly, we constructed a design matrix from the image attributes (using individual attributes, or all attributes), and used multiple linear least squares regression to predict the image-level residual signature. The multiple linear regression was tested using two-fold cross-validation over trials. The relative importance of each attribute (or groups of attributes) was quantified using the proportion of explainable variance (i.e. variance remaining after accounting for noise variance) explained from the residual signature.

3.4.6 Primate behavior zone

In this work, we are primarily concerned with the behavior of an "archetypal human", rather than the behavior of any given individual human subject. We operationally defined this concept as the common behavior over many humans, obtained by pooling together trials from a large number of individual human subjects and treating this human pool as if it were acquired from a single behaving agent. Due to inter-subject variability, we do not expect any given human or monkey subject to be perfectly

consistent with this archetypal human (i.e. we do not expect it to have a human-consistency of 1.0). Given current limitations of monkey psychophysics, we are not yet able to measure the behavior of very large number of monkey subjects at high resolution and consequently cannot directly estimate the human-consistency of the corresponding "archetypal monkey" to the human pool. Rather, we indirectly estimated this value by first measuring human-consistency as a function of number of individual monkey subjects pooled together (n), and extrapolating the human-consistency estimate for pools of very large number of subjects (as n approaches infinity). Extrapolations were done using least squares fitting of an exponential function $\rho(n) = a + be^{-cn}$ (see Figure 3-6).

For each behavioral metric, we defined a "primate zone" as the range of human-consistency values delimited by estimates $\tilde{\rho}_{M_\infty}$ and $\tilde{\rho}_{H_\infty}$ as lower and upper bounds respectively. $\tilde{\rho}_{M_\infty}$ corresponds to the extrapolated estimate of human-consistency of a large (i.e. infinitely many) pool of rhesus macaque monkeys; $\tilde{\rho}_{H_\infty}$ is by definition equal to 1.0. Thus, the primate zone defines a range of human-consistency values that correspond to models that accurately capture the behavior of the human pool, at least as well as an extrapolation of our monkey sample. In this work, we defined this range of human-consistency values as the criterion for success for computational models of primate visual object recognition behavior.

To make a global statistical inference about whether models sampled from the DCNN_{IC} sub-family meet or fall short of this criterion for success, we attempted to reject the hypothesis that, for a given behavioral metric, the human-consistency of DCNN_{IC} models is within the primate zone. To test this hypothesis, we estimated the empirical probability that the distribution of human-consistency values, estimated over different model instances within this family, could produce human-consistency values within the primate zone. Specifically, we estimated a p-value for each behavioral metric using the following procedure: We first estimated an empirical distribution of Fisher-transformed human-consistency values for this model

family (i.e. over all tested DCNN_{IC} models and over all trial-resampling of each DCNN_{IC} model). From this empirical distribution, we fit a Gaussian kernel density function, optimizing the bandwidth parameter to minimize the mean squared error to the empirical distribution. This kernel density function was evaluated to compute a p-value, by computing the cumulative probability of observing a human-consistency value greater than or equal to the criterion of success (i.e. the Fisher transformed $\widetilde{\rho}_{M_\infty}$ value). This p-value indicates the probability that human-consistency values sampled from the observed distribution would fall into the primate zone, with smaller p-values indicating stronger evidence against the hypothesis that the human-consistency of DCNN models is within the primate zone.

3.5 Acknowledgements

This research was performed in collaboration with Elias Issa (equal contributor), Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Reversible inactivation of different millimeter-scale regions of primate IT results in different patterns of core object recognition deficits

Primate core visual object recognition — the ability to rapidly discriminate among objects near the center of gaze in spite of naturally occurring identity-preserving image variability — is thought to rely on the ventral visual stream. Decades of research have shown that IT is a neural correlate of primate recognition behavior, but it is still unclear if and how IT causally supports this general ability. Using reversible inactivation of local regions in IT, we here provide direct causal evidence for the role of IT in object recognition. We found that inactivating different millimeter-scale regions of primate IT resulted in different patterns of object recognition deficits, each predicted by the local region's neuronal selectivity. To the best of our knowledge, this[1] is the first study to demonstrate the necessity of IT cortex for a wide range of general core object recognition behaviors with behaviorally critical topographic organization.

[1]The contents of this chapter are adapted from a journal article in preparation [Rajalingham and DiCarlo, 2018].

4.1 Introduction

Primate core visual object recognition — the ability to rapidly recognize objects in spite of naturally occurring identity-preserving image variability — is thought to rely on the ventral visual stream, a hierarchy of visual cortical areas [DiCarlo et al., 2012]. In particular, decades of research suggest that inferior temporal (IT) cortex, the highest level of the ventral stream hierarchy, is a necessary part of the brain's neural network that underlies core recognition behavior [Logothetis and Sheinberg, 1996, Tanaka, 1996, Rolls, 2000, DiCarlo et al., 2012]. For example, it has been shown that the population of neurons in IT not only matches overall primate behavioral performance [Hung et al., 2005, Zhang et al., 2011] but also predicts primate behavioral patterns [Sheinberg and Logothetis, 1997, de Beeck et al., 2001, Majaj et al., 2015], suggesting that IT is a good neural correlate of primate recognition behavior. These observations are consistent with the causal dependency of core object recognition behavior on IT, but could also reflect epiphenomenal mechanisms [Katz et al., 2016, Liu and Pack, 2017]. For clarity, we adopt the terminology of [Jazayeri and Afraz, 2017], whereby causal dependencies link a dependent variable to an experimentally controlled variable, in contrast to correlational dependencies which are associations that we measure but do not control (e.g. associations between neural activity and behavior measured as visual stimuli are experimentally controlled). Thus, to infer a causal link between activity in IT and behavior, it is necessary to directly manipulate activity in IT (e.g. via the application of pharmacological agents into IT to silence neurons, etc.) while measuring behavior.

To date, the most successful direct manipulations of IT have exclusively targeted at millimeter-scale clusters of face-selective neurons in IT [Afraz et al., 2006, Afraz et al., 2015, Moeller et al., 2017, Sadagopan et al., 2017]. These results suggest that these IT sub-regions are necessary for at least some basic- and subordinate-level face recognition behaviors. However, results from direct manipulations of IT in general visual recognition behavior have been equivocal at best. Lesions of IT sometimes

suggest the necessity of IT and visual behaviors [Cowey and Gross, 1970, Manning, 1972, Holmes and Gross, 1984, Biederman et al., 1997, Buffalo et al., 2000] but the resulting behavioral deficits are often contradictory (with often no lasting visual deficits) [Dean, 1974, Huxlin et al., 2000] and at best modest (e.g. 10-15% drop in performance for large-scale bilateral removal of IT when a complete loss of performance would have been 40%) [Horel et al., 1987, Matsumoto et al., 2016]. Thus, it is still unclear if IT is necessary for general core object recognition behavior. Moreover, even if IT cortex is indeed necessary for all core object recognition tasks, it is unclear if that assumed causal role is spatially organized. For example, the current literature on monkey IT is not inconsistent with the hypothesis that every square millimeter of IT cortex outside of the fMRI-defined face patches is equally involved in all (non-face) object discriminations.

To investigate these open questions, we here reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of IT via local injection of muscimol while monkeys performed a battery of binary core object discrimination tasks, interleaved trial-by-trial. Our results show that inactivation of even single, millimeter-scale regions of IT resulted in reliable contralateral-biased behavioral deficits. Interestingly, these deficits were highly selective over recognition tasks — inactivating a small region of IT produced deficits in only a subset of tasks. Furthermore, inactivating different millimeter-scale regions of primate IT resulted in different patterns of object recognition deficits. Moreover, the effect of inactivation was topographically organized in that the pattern of behavioral deficit was most similar at anatomically neighboring (within 2mm) injection sites. We also found that the pattern of task deficits was well predicted by the local region's neuronal selectivity. Taken together, these results demonstrate the necessity of IT cortex for a wide range of general core object recognition behaviors, and that — even outside of face patches — IT cortex has behaviorally-critical topographic organization for visual features as previously suggested [Wang et al., 1998, Tsunoda et al., 2001].

4.2 Results

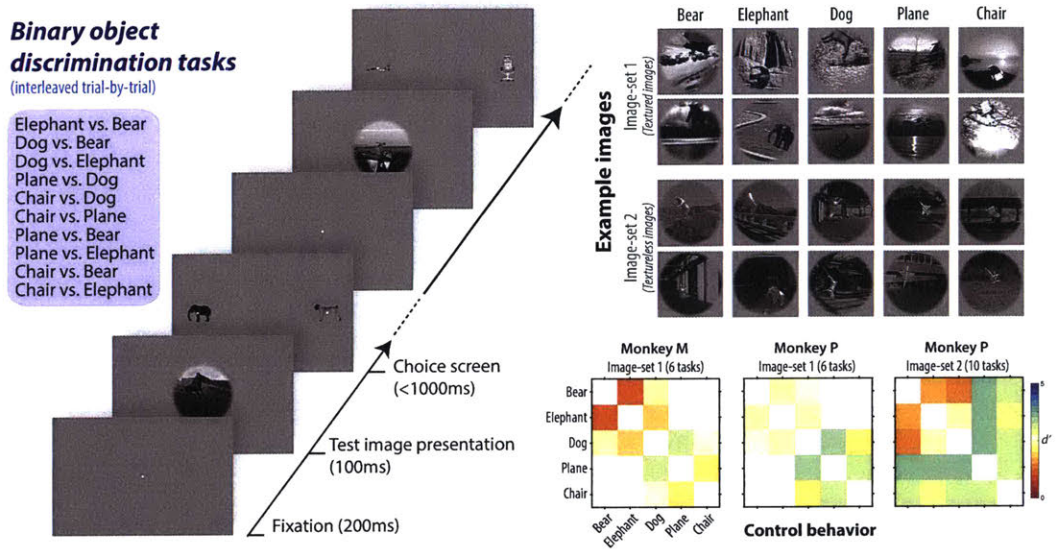

Figure 4-1: : (a) Behavioral paradigm. The list shows all tested pairwise object discrimination tasks between five objects, interleaved trial-by-trial. Each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image (8x8 degrees of visual angle in size) appeared at the center of gaze for 100ms. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right. One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms. Animals were rewarded with small juice rewards for successfully completing each trial. After the end of each trial, another fixation point before the next test image appeared. (b) Visual images. Two (out of hundreds) example images per object, for each of the five objects and for both image sets, are shown. Stimuli consisted of naturalistic synthetic images of 3D objects rendered under high view-uncertainty and overlaid on a naturalistic background. We additionally generated a dataset consisting of texture-less images of the same objects. For the purpose of the current work, we treat both of these image sets as equivalent, namely as images of the same five objects under study. (c) Control behavior. Each matrix shows the control behavioral performance over binary object recognition tasks, for each monkey and image set type. To reliably measure performance for each task within a single behavioral session, we sub-selected six of these ten tasks for most experiments. For a subset of experiments in one animal (monkey P, experiment 2), we tested all 10 binary tasks.

As stated in the introduction, our primary goal was to determine if IT causally supports object recognition, and whether any such causal role is functionally specific at the millimeter-scale. To investigate this, we reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of IT via injection of muscimol while monkeys performed a battery of binary core object discrimination tasks. Figure 4-1 shows the behavioral paradigm used for testing monkeys' core object recognition behavior. We tested several (6 or 10) pairwise object discrimination tasks between five objects, interleaved trial-by-trial (see Figure 4-1 A for task list, and C for control behavior on these tasks). To enforce true invariant recognition, stimuli consisted of naturalistic synthetic images of 3D objects rendered under high view-uncertainty (see 4-1 B for example images).

Figure 4-2A shows the behavioral data for an example inactivation experiment, for each of six tasks. Each panel shows the relative behavioral performance (mean $\pm$ SE, obtained by bootstrap resampling over trials) for a given binary task, for each of three consecutive behavioral sessions (pre-control, inactivation, and post-control; see Methods). Performance is shown relative the average of pre- and post-control performances, which we use as a measure of control behavior (see Methods); the dark and light shaded areas correspond to one and two SE of this measure, respectively. We observed a strong and significant deficit for some tasks (i.e. chair versus dog, chair versus plane, and dog versus bear) but not others (elephant versus bear, dog versus elephant). The resulting pattern of behavioral deficits for this one example inactivation site in IT is shown in Figure 4-2B, with the corresponding anatomical location shown in the inset. Figure 4-2C shows the pattern of behavioral deficits for eight more example inactivation sites in IT from both monkeys (monkey M, P in the first and second row, respectively). We qualitatively observe that inactivating each local region resulted in strong task-specific behavioral deficits. Together, these results suggest that inactivating different millimeter-scale regions of primate IT resulted in different patterns of task deficits. This inference is directly and quantitatively tested in the following analyses.

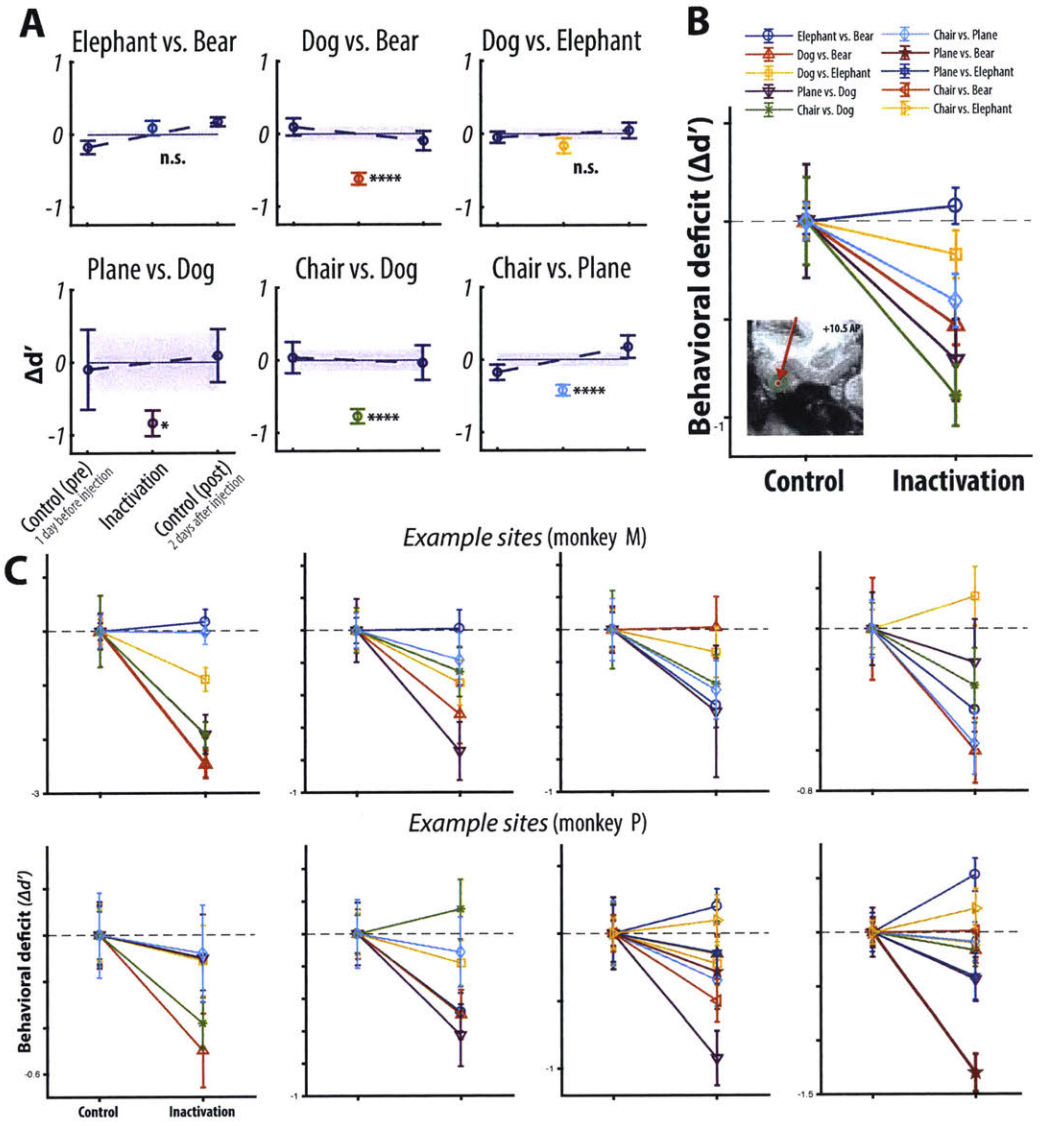

Figure 4-2: (a) Example inactivation experiment. For an example inactivation experiment, the behavioral performance for each of six tasks is shown. Each panel shows the relative behavioral performance (mean ± SE, obtained by bootstrap resampling over trials) for each of three consecutive behavioral sessions (pre-control, inactivation, and post-control; see Methods). Performance is shown relative the average of pre- and post-control performances, which we use as a measure of control behavior (see Methods); the dark and light shaded areas correspond to one and two SE respectively of this measure. We observe a strong and significant deficit for some tasks (i.e. chair versus dog, chair versus plane, and dog versus bear) but not others (elephant versus bear, dog versus elephant). (b) For the example inactivation site in IT in (a), the behavioral deficits are summarized relative to the average control performance on the right panel (mean ± SE over trials). (c) N more example inactivation sites in IT in both monkeys, each with their anatomical locations and resulting behavioral deficits over tasks. Formatting as in 2A.

4.2.1 Summary of behavioral deficits

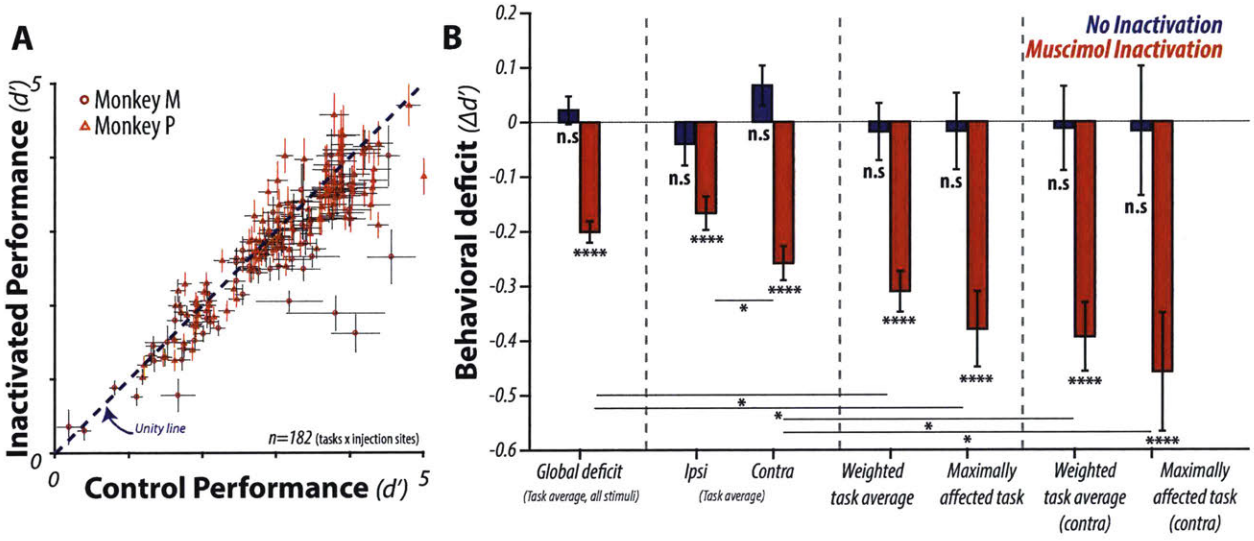

Figure 4-3: : (a) Behavioral deficits for all inactivation sites and all tasks in both monkeys as a scatter of control performance and inactivation performance, showing a significant decrease in performance corresponding to points under the unity line (dashed line). (b) Summary of behavioral deficits. The red bars show the magnitude of inactivation deficit, for all tasks and for all inactivation sites. The blue bars correspond to otherwise identical experiments but without muscimol inactivation. Inactivation of local regions of IT resulted in highly reliable behavioral deficits, which were selective over visual space (i.e. contralateral-biased) and selective over tasks (red bars).

Figure 4-3 shows the behavioral deficits for all inactivation sites and all tasks in both monkeys as a scatter of control performance versus inactivation performance (4-3A). We observed a significant decrease in performance, corresponding to points under the unity line; on average, this amounted to a global deficit of $\mu_\delta = -0.2 \pm 0.02$ in units of d' ($p = 1.23 * 10^{-16}$, one-tailed exact test; see Figure 4-3B, red bar under *global deficit*). We observed no such behavioral deficit on otherwise identical experiments but without inactivation, ($\mu_\delta = 0.02 \pm 0.03$, $p = 0.78$; one-tailed exact test; see blue bar). Consistent with the known lateralization of IT [Op De Beeck and Vogels, 2000], this deficit was more pronounced for contralateral stimuli ($\mu_\delta = -0.26 \pm 0.03$, $p = 1.28e-16$) than for ipsilateral stimuli ($\mu_\delta = -0.17 \pm 0.03$, $p = 3.82 * 10^{-12}$) and

this difference was significant ($p = 0.0128$, one-tailed exact test; ipsi vs. contra). Note that all images were presented foveally (-4 to 4 deg), and contralateral stimuli refers to images where the center of the object was located on the side of the image that corresponds to the contralateral visual hemifield, while potentially still overlapping with both hemifields.

Next we asked whether the inactivation deficits were task-specific. Rather than examine each task individually, we assigned a weight ($w_i \in [0, 1]$) to each task to characterize its deficit, resulting in a weight vector for each inactivation experiment. Weights were obtained by non-negative least squares linear optimization with Tikhonov regularization, enforcing that the sum of all weights over tasks equals one. Intuitively, a weight of zero corresponds to no deficit at all, and a weight of one corresponds to a unique deficit (i.e. only this task was affected). Crucially, the task weights were optimized on held-out data; we split our data into two disjoint halves of trials, optimized the task weights from one split-half, and applied the optimized weight vector on the second split-half (see Methods). Using this procedure, we observed a significantly greater deficit when task are optimally re-weighted ($\mu_\delta = -0.31 \pm 0.04$, $p = 1.63 * 10^{-16}$), indicating that behavioral deficits are not uniform over tasks ($p = 7.09 * 10^{-3}$, one-tailed exact test; weighted task average vs. global). We repeated this procedure with an indicator weight vector, which has a value of one for the most affected task and zero for all others; again, the weight vector was optimized on held-out data (split-half of trials). Applying this re-weighting resulted in the average deficit for the most affected task ($\mu_\delta = -0.38 \pm 0.07$, $p = 1.90 * 10^{-16}$), which was also significantly more pronounced than the global deficit ($p = 9.62 * 10^{-3}$, one-tailed exact test; most affected vs. global). As expected, we observed even larger deficits for the maximally affected task than for the task-weighted average, but did not have sufficient power to distinguish between these two ($p > 0.05$, exact test). Finally, the conjunction of task-selective and contralateral selective effects is shown on the right most bars; we observed even greater deficits for contralateral stimuli when reweighting across tasks ($\mu_\delta = -0.38 \pm 0.06$, $p = 2.12 * 10^{-16}$) or selecting the

maximally affected task ($\mu_\delta = -0.46 \pm 0.10$, $p = 6.31 * 10^{-9}$) than for the global deficit ($p = 0.032, 0.031$ for global contra vs. task-reweight contra and most affected contra, respectively). For each of the analyzed conditions, we observed no significant behavioral deficits on otherwise identical experiments without muscimol inactivation ($p > 0.05$; Figure 4-3B, blue bars). Furthermore, the patterns of deficits across these analyzed conditions were similar for both animals. In summary, inactivation of local regions of IT resulted in highly reliable behavioral deficits, which were selective over visual space (i.e. contralateral-biased) and selective over tasks.

4.2.2 Task-selectivity of deficits

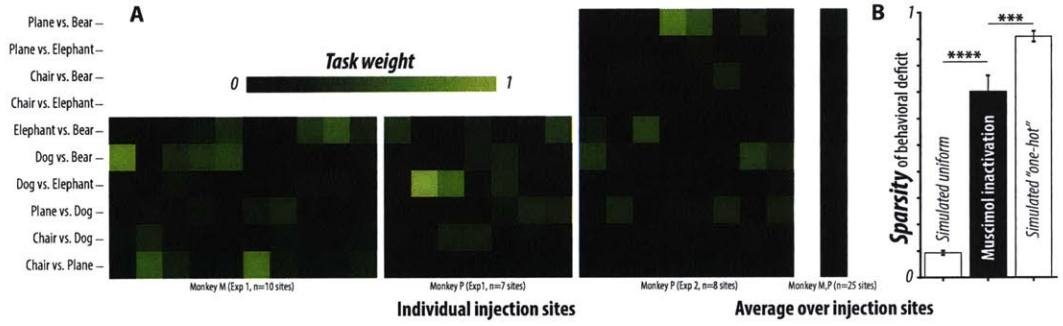

Figure 4-4: (a) The heat map shows the task weight vectors for each of the 25 inactivation sites, with brighter colors corresponding to larger relative task deficits, highlighting that inactivation of different sites resulted in different non-uniform, or relatively sparse, deficit weight pattern. The average weight pattern over all inactivation sites (right column) is largely uniform. (b) Inactivation of local regions in IT leads to significantly non-uniform deficits ($SI = 0.71 \pm 0.05$; mean±SE over sites), as quantified by the sparsity of task weight vectors.

As described above, we characterized each behavioral deficit pattern with a task weight vector w. Figure 4-4 shows the task weight vectors for each inactivation sites as a heat map, where brighter colors correspond to larger relative task deficits for contralateral stimuli. Consistent with the inferred task-selectivity from Figure 4-3, we observed that each inactivation resulted in a non-uniform, or relatively sparse, deficit weight pattern (Figure 4-4A). Importantly, inactivation of different sites led to

different deficit weight patterns (Figure 4-4A, left) while the average weight pattern over all sites is largely uniform (Figure 4-4A, right), indicating that the non-uniformity of task deficits is not tied to specific tasks. Moreover, the task deficit weights were not significantly correlated with task difficulty ($r = 0.06$, $p = 0.39$). Together, these suggest that inactivation of IT results in task-specific behavioral deficits.

We quantified this task-selectivity by computing a sparsity index from each inactivation's behavioral deficit pattern. This index has a value of 0 for uniform deficit patterns, and a value of 1 for a perfectly task-specialized or one-hot deficit pattern. Figure 4-4B shows that inactivation of local regions in IT leads to highly non-uniform deficits ($SI = 0.71 \pm 0.05$; mean±SE over sites); this degree of task selectivity is greater than expected for a uniform deficit ($p = 2.42 * 10^{-16}$; relative to simulated uniform, see Figure 4-4B, Methods) but significantly less than expected for a one-hot deficit pattern ($p = 5.28 * 10^{-3}$; relative to simulated one-hot, see Figure 4-4B, Methods). This inference holds even when computing the sparsity index from a normalized deficit pattern vector ($SI = 0.74 \pm 0.06$; $p = 2.21 * 10^{-6}$, $p = 0.02$ relative to simulated uniform and one-hot, respectively), ensuring again that this non-uniformity does not simply reflect non-uniformity in the behavioral difficulty across tasks. Thus, inactivating local regions in IT results in highly task-selective patterns of behavioral deficits.

4.2.3 Tissue-selectivity of deficits

Figure 4-4A suggests that the patterns of task deficits are also tissue-specific; i.e. inactivating different anatomical regions of IT resulted in different patterns of task deficits. To directly test this, we compared the contralateral deficits of pairs of inactivation; pairwise deficit pattern similarity was quantified using a noise-adjusted correlation ($\tilde{\rho}$, see Methods). We considered all pairs of deficits, measured within the same animal and image-set, that had split-half internal reliability greater than a threshold θ ($n = 62$ pairs for $\theta = 0.1$), but results did not depend on the choice of the threshold θ. We measured the dependence of pairwise deficit similarity on the anatomical distance between the inactivation sites, where anatomical distance

(d) was computed as the Euclidean distance between the injection site locations estimated via high-resolution micro-focal stereo x-ray reconstruction (see Methods). We first observed that inactivation deficits are highly replicable across experiments; the noise-adjusted correlation between behavioral deficit patterns of neighboring inactivation sites was near ceiling ($\tilde{\rho} = 0.92 \pm 0.03$ for $d < 1$mm, Figure 4-5). We further observe that this similarity between the inactivation deficits of two injection sites was monotonically related to the anatomical distance between (Figure 4-5). A simple exponential decay model (half-max-full-width $HMFW = 3.29 \pm 1.19mm$) significantly explained this relationship ($R^2 = 0.36 \pm 0.12$, $p = 8.04 * 10^{-4}$). Given that computing the noise-adjusted correlation required splitting the data into disjoint halves, we did not further split the data (into train and test splits) for cross-validated testing of this model. Instead, we verified that this model correlation is not expected by chance, by fitting the model on randomly shuffled data ($R^2 = 0.00 \pm 0.13$, $p = 0.50$). Together, these results suggest that behavioral deficits are tissue-specific, i.e. the effect of inactivation is different for different inactivation sites, and most similar at anatomically neighboring injection sites.

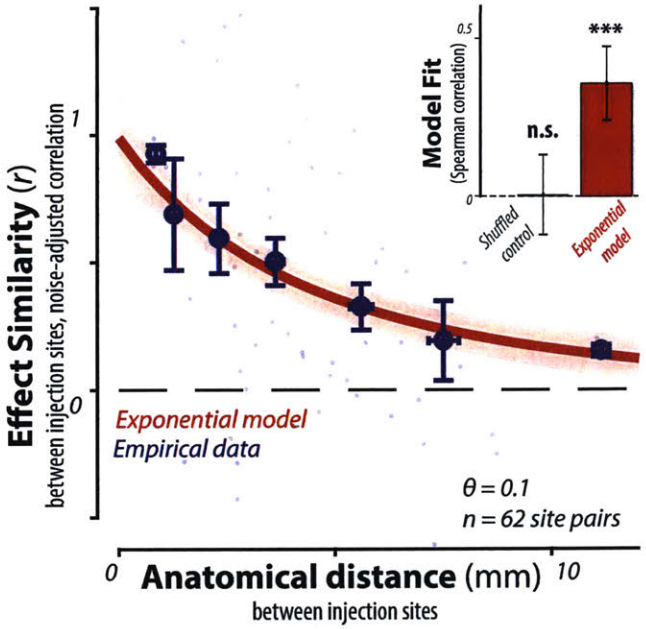

Figure 4-5: (a) Topographical organization. The similarity of behavioral deficit patterns, quantified as a noise-adjusted correlation, between pairs of injection sites is plotted as a function of the anatomical distance between sites. This relationship shows that inactivation deficits are highly replicable; the noise-adjusted correlation between behavioral deficit patterns of neighboring inactivation sites was at ceiling. Moreover, the similarity between any two inactivation deficits was monotonically related to their anatomical distance. Light blue points scatter all pairs. Binned values, with log-spaced sampling of tissue distance, are shown in dark blue (mean ± SE). A simple exponential model significantly explained this relationship (see inset).

4.2.4 Neuronal readout models

Given the observed tissue specificity, we asked to what extent the observed behavioral deficits could be predicted by the local neuronal activity. The central panel in Figure 4-6A shows the location of an example muscimol inactivation site, co-registered with local electrophysiology sites, overlaid on a coronal MRI slice. For this example site in IT, we recorded the activity of eight multi-unit sites (shown in cyan) in close proximity to the injection site (shown in red). Multi-unit activity was recorded in response to the same images as those used in behavioral testing, in a passive viewing paradigm (see Methods). Each sub-panel shows a multi-unit site's stimulus-locked

firing rate responses for each of the five objects, averaged over images. We note that neuronal sites, while heterogeneous, each exhibit reliable object preferences. Based on local neuronal responses such as this, we constructed and tested a number of decoder models, which each map the firing rate image response patterns of local neuronal sites to a predicted behavioral deficit.

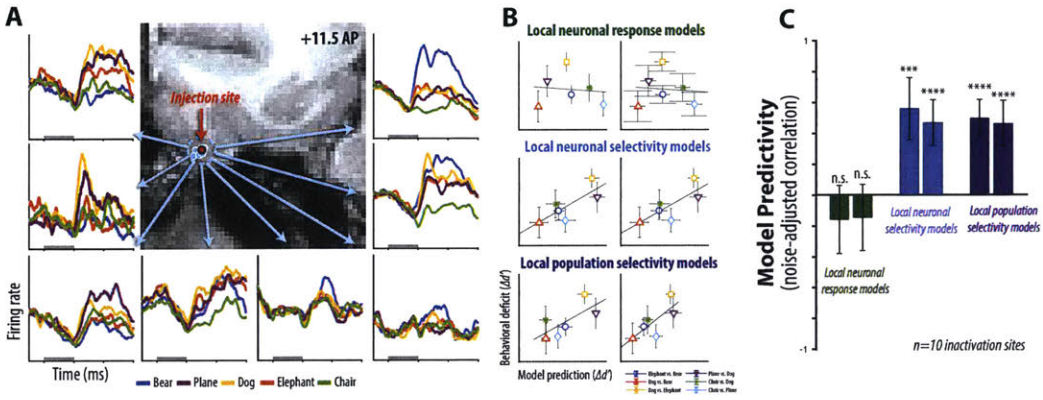

Figure 4-6: (a) Local neurophysiology. For an example muscimol inactivation site, the location of injection co-registered with local electrophysiology recording sites is shown overlaid on a coronal MRI slice. For each the eight neighboring physiology sites, the mean multi-unit visual response aligned to stimulus onset is shown. The stimulus consisted of images of each of the five object categories, and the stimulus duration (0-100ms) is shown with a gray bar. Neuronal sites, while heterogeneous, each exhibit reliable object preferences. (b) To determine whether the observed behavioral deficits are predicted by local neuronal activity, we constructed and tested a number of decoder models that transform these response patterns into predictions of behavioral deficits. The predictions from each of these models, as well as the true (measured) behavioral deficit, are shown for the example inactivation site in (a). The two columns correspond to two variants of decoding models within each class. Note that larger deficits correspond to more negative (i.e. smaller) values of $\Delta d'$. (c) The predictive power of each of these readout models is shown as the noise-adjusted correlation between predicted and actual behavioral deficits, for all relevant sites (with available local physiology on the same images). Of the models that we tested, the most consistent readout model was the local neuronal selectivity.

The tested decoder models roughly correspond to three taxonomical categories: local neural response models, local neural selectivity models, and local population selectivity models. For each category, we tested multiple model instances, varying in the precise details of how neural activity is mapped to the prediction of behavioral

deficits (see Methods). The local neural response models predict largest deficits for tasks with images that yielded largest response from the local neuronal sites. This model class is based on multi-stage readout models (see Methods, Discussion) that predict that neurons that respond highly to particular stimulus classes (e.g. dogs, planes and bears for the example in Figure 4-6A), regardless of whether they explicitly encode differences between them, serve as a domain-specific gates for later discrimination between them (e.g. plane versus dog, or dog versus bear). In contrast, the neural selectivity and population selectivity models predict largest deficits for tasks for which the local neural sample was most discriminative, as measured by a linear classifier. This model class is based on population readout models of IT [Majaj et al., 2015].

We qualitatively observe that the selectivity models better capture the true behavioral deficit pattern than the response models, for this example inactivation site (see 4-6B). This is quantified in Figure 4-6C as a noise-adjusted correlation between predicted and actual behavioral deficits, over all inactivation sites with local neural recordings (n=10 sites for d<1mm). All selectivity models significantly predict the inactivation deficits ($p < 0.001$ for local neuronal and population selectivity models, respectively), while the response models failed to do so ($p > 0.05$ for local response models). In summary, inactivation of millimeter-scale regions of IT results in behavioral deficits that are predicted by the local neuronal activity, and furthermore, the causal link constrains the specific mapping from neurons to behavior.

4.3 Discussion

In this work, we sought to investigate if and how neural activity in IT causally supports core object recognition behavior. To answer this, we reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of IT while monkeys performed a battery of object discrimination tasks. The conceptual advance of this work is two-fold. First, we provide direct causal evidence for the role of IT in core object recognition, which was largely lacking — especially beyond the specific case of face-

selective sub-regions of IT. Second, we uncovered that the causal role of IT in object recognition is topographically organized and predicted by the local neuronal selectivity. These phenomena —namely, the magnitude and sparsity of behavioral deficits from millimeter-scale inactivations, their pairwise similarity as a function of anatomical distance, and the consistency to some but not all neural decoding models— are strong constraints for computational models of the ventral stream and its role in core object recognition behavior.

4.3.1 Direct causal evidence for the role of IT in core object recognition

To fix terminology, we first define the following decoding hypothesis: IT cortex is a necessary part of the brain's neural network that underlies core recognition behavior — or, stated in other words, core object recognition behavior causally depends on IT cortex. To test this decoding hypothesis, we adopt the terminology of [Jazayeri and Afraz, 2017], whereby causal dependencies can be inferred by linking a dependent variable to an experimentally controlled variable, in contrast to correlational dependencies which are associations between variables that are measured but not experimentally controlled. Thus, to infer a causal link between activity in IT and behavior, it is necessary to directly manipulate activity in IT (e.g. via the application of pharmacological agents into IT to silence neurons, etc.) while measuring behavior. Related correlational dependencies (e.g. via direct manipulation of visual input to the retinae while measuring variations from both IT activity and behavior) are consistent with causal dependencies but could also reflect epiphenomenal mechanisms; i.e. correlation does not imply causation. Recently, research in other behavioral domains has exposed potential epiphenomenal mechanisms [Katz et al., 2016, Liu and Pack, 2017], highlighting the need to test directly causal dependencies.

With respect to our stated decoding hypothesis, decades of neurophysiological and neuropsychological research have uncovered correlational dependencies between activity in IT cortex and primate object recognition behavior [Logothetis and Shein-

berg, 1996, Tanaka, 1996, Rolls, 2000, DiCarlo et al., 2012]. Individual neurons in IT cortex are selective to complex visual features in images, and exhibit remarkable tolerance to changes in viewing parameters [Kobatake and Tanaka, 1994, Ito et al., 1995, Logothetis et al., 1995, Booth and Rolls, 1998, Rust and DiCarlo, 2010]. Moreover, a simple readout from the population of neurons in IT not only matches overall primate behavioral performance [Hung et al., 2005, Zhang et al., 2011] but also reliably predicts the behavioral error patterns [Majaj et al., 2015]. Taken together, these results are consistent with our decoding hypothesis, but could also reflect epiphenomenal mechanisms. Direct causal evidence is still largely lacking for this hypothesis. To this end, our first major contribution in this work is to provide direct causal evidence for the role of IT in core object recognition behavior.

Prior to this, causal evidence for the role of IT in core object recognition has been both scarce and equivocal. Lesions of IT suggest a coarse causal link between this area and visual behaviors [Cowey and Gross, 1970, Manning, 1972, Holmes and Gross, 1984, Weiskrantz and Saunders, 1984, Buffalo et al., 1998, Huxlin et al., 2000, Matsumoto et al., 2016] but the resulting behavioral deficits are often contradictory [Dean, 1974, Huxlin et al., 2000] and at best modest [Horel et al., 1987, Matsumoto et al., 2016]. For example, recent work showed that near complete ablation of IT (bilateral removal of anterior IT) resulted in only mild (10-15%) deficits in object categorization [Matsumoto et al., 2016]. Similar modest behavioral deficits on visual recognition tasks were also observed with large-scale reversible inactivation via cooling of the temporal lobe [Horel et al., 1987]. It is unclear to what extent these modest behavioral deficits can be explained by limitations of the methodologies and the behavioral assays, which may not be robust to alternative (potentially compensatory) behavioral strategies. Several other (higher-resolution) methodologies using focal reversible neural perturbation methods (e.g. electrical, pharmacological, optogenetic and chemogenetic perturbations) have been successfully used in testing decoding hypotheses in other behavioral domains [Salzman et al., 1990, Recanzone et al., 1992, Celebrini and Newsome, 1995, Britten and van Wezel, 1998, DeAngelis et al., 1998, Romo et al., 1998, Thier and Andersen, 1998, Romo et al., 2000, Bisley et al., 2001, Nichols and

Newsome, 2002, Zhang et al., 2011, Jazayeri et al., 2012, Dai et al., 2014, Eldridge et al., 2016] (we note that these methods likely do not completely rule out the possibility of dynamic downstream compensation [Fetsch et al., 2018]). Only a handful of studies have reported using focal reversible neural perturbation tools to test the stated decoding hypothesis in IT. Interestingly, these studies exclusively targeted spatial clusters of face-selective neurons in IT, testing the causal role of these regions in basic- and subordinate-level face recognition behaviors [Afraz et al., 2006, Afraz et al., 2015, Moeller et al., 2017, Sadagopan et al., 2017] (beyond object recognition, one study tested the causal role of spatial clusters of disparity selective neurons in a disparity discrimination task [Verhoef et al., 2012]). Thus, our results provide much needed direct causal evidence for the general decoding hypothesis.

4.3.2 The causal role of IT in object recognition is topographically organized

While faces are an especially behaviorally relevant stimulus class for primates [Tsao and Livingstone, 2008], the experimental bias towards face-selective spatial clusters in IT is likely related to the spatial resolution limitations of current neural perturbation tools, which operate on groups of spatially contiguous neurons at approximately millimeter-scale. Given this limitation, the known millimeter-scale spatial clusters of face selective regions in IT [Tsao et al., 2003, Tsao et al., 2006, Tsao and Livingstone, 2008] form an intuitively optimal candidate for testing causal dependencies related to our decoding hypothesis. We note that similar spatial clustering has been reported for a small number of stimulus domains [Conway et al., 2007, Kornblith et al., 2013, Lafer-Sousa and Conway, 2013, Verhoef et al., 2015]. Given that these regions respond preferentially to images of faces over other objects, previous studies targeting these regions have tested their causal role in basic- and subordinate-level face recognition behaviors [Afraz et al., 2006, Afraz et al., 2015, Moeller et al., 2017, Sadagopan et al., 2017]. Importantly, the topographic organization of neurons in IT is largely unknown and assumed by many to be functionally random and heterogeneous be-

yond these discrete clusters. To support a general inference, we here tested arbitrary sampled millimeter-scale regions of ventral IT, rather than functionally target inactivation sites. Interestingly, we found that inactivation of different regions in ventral IT led to different task-specific deficits, suggesting some functional specificity for arbitrarily sampled millimeter-scale regions. Based on our data, it is unclear whether this topographic organization is stereotyped, as is the case with previously reported discrete clusters [Tsao et al., 2006], or highly variable across different subject. This topographical organization is consistent with previously reported sub-millimeter scale columnar organization of neurons in IT [Fujita et al., 1992, Tanaka, 1996, Wang et al., 1996, Wang et al., 1998]. We speculate that this topographic organization reflects a general principle of global cortical layout, whereby neuronal selectivities are optimized in the face of metabolic constraints (e.g. minimization of connection wiring length [Chklovskii et al., 2002]). These phenomena could guide computational models of the ventral stream and its role in core object recognition behavior.

4.3.3 The causal role of IT in object recognition is predicted by the local neuronal selectivity

Finally, we found that behavioral deficits from inactivating millimeter scale regions of IT are consistent with predictions from a spatially distributed readout of neurons in IT, as evidenced by the ability of particular local neural selectivity readout models to predict inactivation deficits. In contrast, inactivation deficits were not well predicted by particular local neural response readout models. These models are one possible instantiation of a class of multi-stage readout models that frame detection and discrimination as separate, sequential stages; such models have been proposed for a number of putative specialized domains [Tsao and Livingstone, 2008, Chang and Tsao, 2017]. Multi-stage readout models predict that neurons that respond highly to particular stimulus classes (e.g. dogs and planes), without explicitly encoding the differences between them, serve as domain-specific gates for later discrimination between them. Thus, one might expect that inactivating these neurons should result in large

behavioral deficits for such discriminations (i.e. for discriminating between dogs and planes). Previously, it was difficult to discriminate between these two model classes, as detection and discrimination ability were highly correlated within face-selective neurons [Afraz et al., 2015], but our data here are sufficiently powerful to make this distinction. In summary, our data are consistent with at least one readout model, and provide constraints for discriminating between alternative readout models that link neural responses to object recognition behaviors.

4.4 Methods

4.4.1 Subjects and surgery

Two adult male rhesus macaque monkeys (Macaca mulatta, subjects M, P) were trained on the core object recognition paradigm described below. For each animal, a surgery using sterile technique was performed under general anaesthesia to implant a titanium head post to the skull using titanium screws, and a steel cylindrical recording chamber (19 mm inner diameter; Crist Instruments) over a craniotomy targeting the temporal lobe in the left hemisphere from the top of the skull (Monkey M, +13 mm posterior-anterior, +16.3 mm medial-lateral, 15° medial-lateral angle; Monkey P, +13 mm posterior-anterior, +14.75 mm medial-lateral, 15° medial-lateral angle). All procedures were performed in compliance with National Institutes of Health guidelines and the standards of the MIT Committee on Animal Care and the American Physiological Society.

4.4.2 Core object recognition behavioral paradigm

Core object discrimination is defined as the ability to discriminate between two or more objects in visual images presented under high view uncertainty in the central visual field ($\sim 10°$), for durations that approximate the typical primate, free-viewing fixation duration ($\sim$ 200 ms) [DiCarlo and Cox, 2007, DiCarlo et al., 2012]. As in our previous work [Rajalingham et al., 2015, Rajalingham et al., 2018], we investigate

this behavior using an interleaved set of binary match-to-sample discrimination tasks. The behavioral paradigm is described below. Behavioral data was collected under head fixation, and subjects reported their choices using their gaze. We monitored eye position by tracking the position of the pupil using a camera-based system (SR Research Eyelink 1000). Images were presented on a 27" LCD monitor (1920 x 1080 at 60 Hz; Samsung S27A850D) positioned 44 cm in front of the animal. At the start of each training session, subjects performed an eye-tracking calibration task by saccading to a range of spatial targets and maintaining fixation for 800ms. Calibration was repeated if drift was noticed over the course of the session.

Figure 4-1A illustrates the behavioral paradigm. Each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image ($8 \times 8°$ of visual angle in size) appeared at the center of gaze for 100ms. Trials were aborted if gaze was not held within $\pm 2°$. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right (each centered at 8° of eccentricity along the horizontal meridian; see Fig. 1B). One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The object that was not displayed in the test image is referred to as the distractor object, but note that objects are equally likely to be distractors and targets. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms. The monkey was rewarded with small juice rewards for successfully completing each trial. After the end of each trial, another fixation point appeared, cueing the next trial. Each trial consisted of a different randomly selected binary task. Real-time experiments for monkey psychophysics were controlled by open-source software (MWorks Project http://mworks-project.org/).

Both animals were previously trained on other images of other objects, and were proficient in discriminating between over 35 object categories. In this study, five new objects were tested, which resulted in ten possible binary object discrimination tasks

(see Figure 4-1A for complete list). To reliably measure performance for each task within a single behavioral session, we sub-selected six of these ten tasks for most experiments. For a subset of experiments in one animal (monkey P, experiment 2), we tested all 10 binary tasks. For each session, monkeys were tested for several hours (until satiation) and performed a large number of trials (monkey M: 3442±1097, monkey P: 4430±942; mean ± SD). Figure 4-1C shows the control behavioral performance for each monkey and image set.

4.4.3 Visual images

We examined basic-level object recognition behavior using naturalistic synthetic images of a set of five objects. The image generation pipeline is described in detail elsewhere. Briefly, each image was generated by first rendering a 3D model of the object with randomly chosen viewing parameters (2D position, 3D rotation and viewing distance), and then placing that foreground object view onto a randomly chosen, natural image. Object models spanned basic-level object categories (bear, elephant, dog, airplane, and chair). Background images were sampled randomly from a large database of high-dynamic range images of indoor and outdoor scenes obtained from Dosch Design (www.doschdesign.com). This image generation procedure enforces invariant object recognition, rather than image matching, as it requires the animal to tackle the invariance problem, the computational crux of object recognition [Ullman and Humphreys, 1996, Pinto et al., 2008].

The majority of the behavioral data presented here was collected in response to a base image set of five objects rendered with the image generation pipeline described above (40 images/object, 200 images in total). We additionally generated a variant of this dataset consisting of texture-less images of the same objects. These images were targeted to both titrate the task difficulty and further remove potential low-level confounds (e.g. luminance and contrast). For this image set, new images of the same objects with the same generative parameters were generated on each behavioral session, while holding a portion of images (20%) fixed across sessions. For the purpose of the current work, we treat both of these image sets as equivalent, namely as

images of the same five objects under study differing only in their precise generative parameters. Figure 4-1B shows example two images for each object, from both image sets.

4.4.4 Physiology and pharmacology

In each animal, we first recorded multi-unit activity (MUA) from randomly sampled sites on the ventral surface of IT (monkey M: 57 multi-unit sites, monkey P: 43 multi-unit sites). Recordings were made using glass-coated tungsten microelectrodes (impedance, $0.3-0.5M\Omega$; outer diameter, 310um; Alpha Omega). A motorized microdrive (Alpha Omega) was used to lower electrodes through a 26-gauge stainless-steel guide tube inserted into the brain (5 mm) and held by a plastic grid inside the recording chamber (CRIST). We recorded MUA responses from IT while monkeys passively fixated images in a rapid serial visual presentation (RSVP) protocol (10 images/trial, 100ms on, 100 ms off). To ensure accurate stimulus presentation, eye position was tracked and trials were aborted if gaze was not held within $\pm 1.5°$. To ensure accurate stimulus locking, spikes were aligned to a photodiode trigger attached to the display screen. Multi-unit responses were amplified (1x head-stage), filtered (250Hz cutoff), digitized (sampling rate of 40kHz) and sorted (Plexon MAP system, Plexon Inc.). Firing rates were computed as the total number of spikes in two post-stimulus windows (70-170ms, 170-270ms).

Following this mapping stage, we performed inactivation experiments using focal microinjections of muscimol, a potent GABA agonist [Andrews and Johnston, 1979]. We varied the location of microinjections to randomly sample the ventral surface of IT. Given the relatively long half-life of muscimol, inactivation sessions were interleaved over days with control behavioral sessions. Each inactivation experiment consisted of three behavioral sessions: the baseline or pre-control session (1 day prior to injection), the inactivation session, and the recovery or post-control session (2 days after injection). Each inactivation session began with a single focal microinjection of 1ul of muscimol (5mg/ml, Sigma Aldrich) at a slow rate (100nl/min) via a 30-gauge stainless-steel cannula at the targeted site in ventral IT. Injections were made with

through a simple microinjection circuit consisting of a three-way valve (Labsmith) and marker line (similar to [Noudoost and Moore, 2011]), enabling precise monitoring of the flow and volume of muscimol injected. In pilot experiments, we verified complete neural suppression at the location of injection using custom-built single-use injectrodes [Noudoost and Moore, 2011]. Given this volume of muscimol, we estimate strong neural suppression within a local region of 2.5mm in diameter (with partial suppression within a 4mm diameter region) for up to six hours after injection [Arikan et al., 2002]. Immediately after injection, we waited 10-20 minutes before measuring the monkey's behavior on a battery of object recognition tasks for up to 3 hours post-injection.

To ensure accurate targeting of IT, all electrophysiological recordings and pharmacological injections were made under micro-focal stereo x-ray guidance [Cox et al., 2008]. Briefly, monkeys were fitted with a plastic frame (3 x 4 cm) positioned near the temporal lobe using a plastic arm anchored in the dental acrylic implant. The frame contained six brass fiducial markers (1mm diameter) of known geometry, measured using micro-CT. The fiducial markers formed a fixed 3D skull-based coordinate system for registering all physiological recordings and pharmacological injection sites. At each site, two x-rays were taken simultaneously at near orthogonal angles, and the 3D location of the electrode/cannula tip was reconstructed relative to the skull using stereo-photogrammetric techniques. This procedure enables high-resolution reconstruction (<200um error) of electrode and cannula locations across experimental sessions [Cox et al., 2008, Issa et al., 2013].

In total, we collected data for 25 inactivation experiments in two monkeys (monkey M: $n = 10$ experiments, monkey P: $n = 15$ experiments). Throughout the experimental data collection period, we additionally interleaved control experiments of three consecutive control behavioral sessions each, with the same images and tasks but with no injections. These data ($n = 18$ experiments), matching the three-session design of inactivation experiments, form a control condition against natural inter-session variability.

4.4.5 Analysis

Behavioral metrics

We previously introduced several metrics to characterize behavior in this binary match-to-sample paradigm [Rajalingham et al., 2018]. Here, we focus on the highest resolution behavioral metric that can be reliably measured in a single behavioral session, the one-versus-other object level performance metric (termed B.O2). Briefly, this metric is a pattern of pairwise object discrimination performances. For each pairwise object discrimination task, performance was estimated using a sensitivity index d' [Macmillan, 1993]: d' = Z(hit rate) - Z(false alarm rate), where Z(.) is the inverse of the cumulative Gaussian distribution. All d' estimates were constrained to a range of [0,5].

Recall that each inactivation experiment consisted of three behavioral sessions. We first equated the number of trials per session by selecting the first N trials of each session, where N was the minimum number of trials across the three sessions. For each of these three behavioral sessions, we then computed a pattern of performances across tasks (b). To measure the behavioral deficit from inactivation, we estimated a behavioral deficit pattern (δ) as the difference between inactivated and control performance over tasks: $\delta = \psi_{inactivated} - \psi_{control}$. The control behavioral performance was defined as the average of the pre-control and post-control performances: $\psi_{control} = (\psi_{precontrol} + \psi_{postcontrol})/2$.

Task weight

For each behavioral deficit pattern δ, we estimated a task weight vector w to characterize the deficit task-selectivity. To do so, we first calculated a matrix D where each row is an estimate of δ obtained via bootstrap resampling of trials. w was obtained by non-negative least squares optimization with Tikhonov (or ridge) regularization to minimize the weighted deficit $\|D \cdot w\|$. We then enforced that weights sum to 1

over tasks by normalizing by the sum:

$$\tilde{w} = \text{argmin}_{w_i \geq 0}(-\|D \cdot w\| + \lambda^2 \|w\|)$$

$$W = \tilde{W}/\|\tilde{W}\|$$

Intuitively, a task weight (w_i) of zero corresponds to no deficit on task i, and a weight of one corresponds to a unique deficit (i.e. only task i was affected). Task weight vectors were used to re-weight the task deficit pattern to compute a weighted average deficit, rather than global deficit. Crucially, task weights were optimized on held-out data to avoid double-dipping [Kriegeskorte et al., 2009]. To do, we split our data into two disjoint halves of trials, optimized the task weights from the behavioral deficit patterns estimated from one split-half, and applied the optimized weight vector on the second split-half.

We repeated this procedure with an indicator weight vector, which has a value of one for the most affected task and zero for all others. Applying this weight results in the average deficit for the most affected task. Again, the weight vector was optimized on held-out data (split-half of trials).

Sparsity of deficit

We quantified the non-uniformity of the behavioral deficits using a sparsity index $SI(x)$ [Vinje and Gallant, 2000] as follows:

$$A(x) = E[x]^2/(E[x^2]),$$

$$SI(x) = (1 - A(x))/(1 - 1/N)$$

When applied to a behavioral deficit pattern, $SI(\delta)$, this index has a value of 0 for uniform deficit patterns, and a value of 1 for a one-hot deficit pattern. To ensure that the sparsity of the behavioral deficit did not purely reflect non-uniformity in the behavioral difficulty across tasks, we additionally computed this index from a normalized deficit pattern vector: $\delta = \frac{\psi_{inactivated} - \psi_{control}}{\psi_{inactivated} + \psi_{control}}$.

We compared the resulting SI estimate to those expected by simulated uniform and one-hot deficit patterns, respectively. For each site, we first obtained a matrix of deficit pattern estimates, with rows corresponding to different estimates obtained by bootstrap resampling over trials, and columns corresponding to different tasks. To simulate the uniform deficit pattern, we shuffled this matrix (across both dimensions), thus removing any task specific structure while ensuring that the global deficit was left unchanged. To simulate the one-hot deficit pattern, we replaced each row of this matrix with a one-hot pattern, by setting all but the minimum value to zero. In all cases, we first averaged this matrix over bootstrap estimates before computing the SI. By estimating one SI value for each injection site, we obtained a distribution over injection sites.

Neuronal readout models

To investigate the link between neuronal activity and inactivation deficits, we constructed and tested a number of decoding models. Each of these models predicts an inactivation pattern from the activity of neurons recorded in close anatomical proximity (within 2mm) to the injection site. As described above, multi-unit neuronal activity was measured in response to the same images under a passive viewing paradigm. For each recorded multi-unit site, we estimated a vector of firing rate responses over images. The tested decoder models, which map these firing rate response vectors to behavioral deficit predictions, roughly correspond to three taxonomical categories: local neural response models, local neural selectivity models, and local population selectivity models. For each category, we tested two model instances, varying in the precise details of how neural activity is mapped to the prediction of behavioral deficits, described in detail below.

The local neural response models predict largest deficits for tasks with images that yielded largest response from the local neuronal sites. In practice, task deficit predictions were computed as the negative of the average response for images in the task, averaged over all multi-unit sites in the local population (i.e. all sites within 2mm of the inactivation site). Variants of this model class differ in how neural response

vectors were normalized (z-score versus midrange normalization) prior to averaging across units. This model class is inspired from multi-stage readout models that predict that neurons that respond highly to particular stimulus classes, regardless of whether they explicitly encode differences between them, serve as domain-specific gates for later discrimination between them [Tsao and Livingstone, 2008].

In contrast, the local neural selectivity and local population selectivity models predict largest deficits for tasks for which the local neural population was most discriminative, as measured by a linear classifier. We implemented the local neural selectivity model predictions by training linear classifiers (binary linear SVMs) to discriminate between objects from the mean neural response vector, averaged over units. Again, variants of this model class differ in how neural response vectors were normalized (z-score versus midrange normalization) prior to averaging across units. We additionally implemented the local population selectivity model predictions by training linear classifiers (binary linear SVMs) from the entire population of local neuronal sites. In the first variant, we independently mapped each unit's response vector to a behavioral deficit (as in the local neural selectivity model), and estimated the average behavioral deficit prediction. In the second variant, we concatenated the response vectors of all units in the local population to construct a feature matrix, and trained linear classifiers on this matrix. For each of these four models, the predicted deficit for a binary task was estimated as the negative of the cross-validated binary discrimination performance for that task, in units of d'. The local neural selectivity and local population selectivity models were loosely inspired from population readout models of IT [Majaj et al., 2015]. Note, however, that the current implementations do not include the remaining (non-local) IT population as inputs, as we did not have access to a larger sample of IT.

Noise-adjusted correlations

We measured the similarity between two behavioral deficit patterns δ_1, δ_2 (e.g. between true deficit patterns and predictions from a model) using a noise-adjusted correlation [DiCarlo and Johnson, 1999, Johnson et al., 2002]. For each behavioral

deficit pattern, we split all independent raw observations (e.g. behavioral trials) into two equal halves and computed the behavioral deficit pattern from each half, resulting in two independent estimates of the deficit pattern. We took the Pearson correlation between these two estimates as a measure of the reliability of that behavioral deficit pattern, given the data, i.e. the split-half internal reliability. To estimate the noise-adjusted correlation between two deficit patterns, we compute the Pearson correlation over all the independent estimates of deficits from each, and we then divide that raw Pearson correlation by the geometric mean of the split-half internal reliability of each deficit:

$$\widetilde{\rho}(\delta_1, \delta_2) = \frac{\rho_{\delta_1,\delta_2}}{\sqrt{\rho_{\delta_1,\delta_1} \times \rho_{\delta_2,\delta_2}}}$$

Since all correlations in the numerator and denominator were computed using the same amount of trial data (exactly half of the trial data), we did not need to make use of any prediction formulas (e.g. extrapolation to larger number of trials using Spearman-Brown prediction formula). This procedure was repeated 10 times with different random split-halves of trials. Our rationale for using a reliability-adjusted correlation measure was to account for variance in the behavioral deficit that is not replicable by the task condition. If two behavioral deficits are identical, then their expected noise-adjusted correlation is 1.0, regardless of the finite amount of data that are collected. The noise-adjusted correlation was used to compute the similarity between observed and predicted behavioral deficit patterns (e.g. for testing neural readout models), as well as for the similarity between two different behavioral deficit patterns arising from two different inactivation sites.

Statistical testing

Unless otherwise specified, we estimated the uncertainty in delta measurements via bootstrap resampling of trials, repeated 100 times. The standard error of delta measurements was estimated as the standard deviation of this bootstrap distribution. For statistical tests, we performed one-tailed exact tests, by computing the empirical probability of observing a sample below zero. To compute this probability from

the empirical bootstrap distribution, we fit a Gaussian kernel density function to the empirical distribution, optimizing the bandwidth parameter to minimize the mean squared error (kde.m on MATLAB file exchange). This kernel density function was evaluated to compute a p-value, by computing the cumulative probability of observing a positive behavioral delta.

4.5 Acknowledgements

This research was performed in collaboration with James J. DiCarlo.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Towards a chronically implantable LED arrays for optogenetic experiments in primates.

In this work, we sought to develop a novel neural perturbation tool to increase the scale and throughput of current neural perturbation experiments in primates. To this end, we tested a chronically implantable LED array for optogenetic perturbation in primates in two different experiments, each testing the causal role a visual cortical area in an established behavioral task. Our data do not support strong inferences about the utility of this tool, in its current form, for neural perturbation experiments in primates. Rather, these results provide a report of preliminary findings with guides for improvements, both technological and experimental.

5.1 Introduction

Neural perturbation tools — pharmacological, electrical and optogenetic — have helped establish detailed causal links between neural activity and a behavior of interest for many behavioral domains [Salzman et al., 1990, Recanzone et al., 1992, Celebrini and Newsome, 1995, Britten and van Wezel, 1998, DeAngelis et al., 1998, Romo et al., 1998, Thier and Andersen, 1998, Romo et al., 2000, Bisley et al., 2001, Nichols

and Newsome, 2002, Zhang et al., 2011, Jazayeri et al., 2012, Dai et al., 2014, Afraz et al., 2006, Afraz et al., 2015, Moeller et al., 2017, Sadagopan et al., 2017, Verhoef et al., 2012]. Moreover, they are necessary in order to directly test decoding hypotheses, i.e. to infer causal dependencies between neural activity and behaviors with high confidence [Jazayeri and Afraz, 2017]. In particular, optogenetic perturbations, whereby light-sensitive ion channels are embedded in the membrane of genetically targeted neurons in order to modulate their activity via delivery of light [Deisseroth, 2011], has shown remarkable promise in many systems neuroscience applications. The key advantage of optogenetic silencing over other perturbation tools is the ability to inhibit neural activity with precise temporal delimitation (10-200ms) and cell-type specificity [Han et al., 2011, Deisseroth, 2011]. However, the optogenetic toolbox is still relatively under-developed for the primate compared to the rodent model, with only a handful of studies showing behavioral effects of optogenetic perturbation, across diverse behavioral domains [Gerits et al., 2012, Jazayeri et al., 2012, Cavanaugh et al., 2012, Ohayon et al., 2013, Dai et al., 2014, Afraz et al., 2015, Fetsch et al., 2018].

In this work, we sought to improve the utility of optogenetic perturbations in primates by modifying the method of light delivery. We tested a novel, custom developed (Blackrock Microsystems), chronically implantable array of LEDs for optogenetic experimentation in primates. Our primary motivation was to develop a chronic perturbation tool, as this could enable experimental measurements over weeks and months, thus increasing the scale and throughput of current causal experiments. For example, current perturbation experiments are limited to inferences over collections of images (e.g. several images of the same object), for a small set of such collections; a chronic tool, if successful, could enable the collection of large-scale datasets at individual image resolution, for thousands of images. In addition to this main motivation, this particular tool offers the promise of precise spatial and temporal control for neural perturbation. In this regard, time-delimited perturbation at the resolution of tens of milliseconds could be useful for examining the dynamics of the neural code. Analogously, large spatial tiling at fine resolution could enable tackling questions of topographic organization of neural codes.

To this end, we here applied this novel tool in two experimental studies, each testing the causal role of a visual cortical area in an established behavioral task. Our first experimental condition tested the causal role of primary visual (V1) cortex in a simple luminance discrimination task, while the second experimental condition tested the causal role of interior temporal (IT) cortex in core object recognition. Our preliminary results suggest that, in each case, neural suppression using this tool may result in reliable behavioral effects, at least transiently. However, in its current implementation, this chronic tool is far from reliable or high-throughput, as compared to alternative methods. We suggest possible technological improvements that may increase its utility for our systems neuroscience goals.

5.2 Results

As stated in the Introduction, we aimed to test a novel chronically implantable LED array for optogenetic experiments in primates, shown in Figure 5-1A. Briefly, each LED array consists of a 5x5 printed circuit board grid with 24 LEDs and one thermal sensor for monitoring tissue heating from light power. Each LED is 0.5mmx0.5mm, with 1mm spacing between LEDs. The PCB and LEDs are encapsulated within a thin ($<$ 0.5mm) translucent silicone cover. The LED array is designed to be chronically implanted directly on the cortical surface, by suturing the silicone encapsulation onto the dura mater.

We first measured the photometric properties of the LED array for direct comparison with an alternative light delivery method for optogenetic perturbation (optrodes, consisting of an optic fiber, coupled to LASER, that is acutely inserted into the cortical tissue). Figure 5-1B shows the total light power output of a given LED, as a function of applied voltage plotted as percentage of the maximum voltage. Measurements were made with a power-meter in tight proximity ($<$ 0.5mm) to the LED arrays, averaging the power output over a detector of 9mm in diameter and over a 500ms duration window. We note that an individual LED operating at 30% can match the power output of optrodes that have successfully yielded behavioral effects

in monkeys ($\sim 10 - 15$mW). Naturally, much more power can be delivered via an optrode by optimizing the LASER coupling.

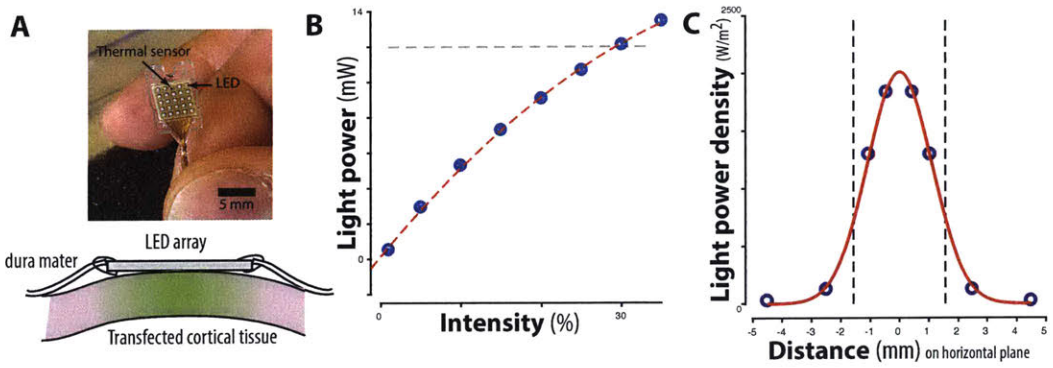

Figure 5-1: (a) The top panel shows a photograph of LED array, a 5x5 grid with 24 LEDs and one thermal sensor. The LED array is designed to be chronically implanted directly onto the cortical tissue, by suturing the thin silicone encapsulation onto the dura mater, as illustrated in the bottom panel. (b) Light power output for individual LEDs as a function of the input intensity (controlled via input voltage). The horizontal line corresponds to average power output of optrodes that have successfully yielded behavioral effects in monkeys. (c) Spatial density of light power on the horizontal plane, at a transverse distance of $<$ 1mm from the surface of the LED. Given that light delivered from LEDs is not collimated, the spatial spread of light power over the horizontal plane is relatively large ($\sim$ 2.5mm)

However, this power output is delivered in a spatially non-uniform manner. Figure 5-1C shows the spatial density of light power on the horizontal plane, at a transverse distance of $<$ 1mm from the surface of the LED. Even without accounting for tissue absorption of light, the measured values are several orders of magnitude smaller than the corresponding *irradiance* estimates for acute optrodes in cortical tissue [Chow et al., 2010]. Given that light delivered from LEDs is not collimated, the spatial spread of light power over the horizontal plane is relatively large ($\sim$ 2.5mm). Thus, while individual LEDs are sub-millimeter in size (0.5mm x 0.5mm), the effective spatial resolution of this tool is several millimeters. Together, these results suggests that the total light power to impinge on individual spatially targeted neurons is significantly lower than for current acute light delivery methods.

5.2.1 Perturbation of V1

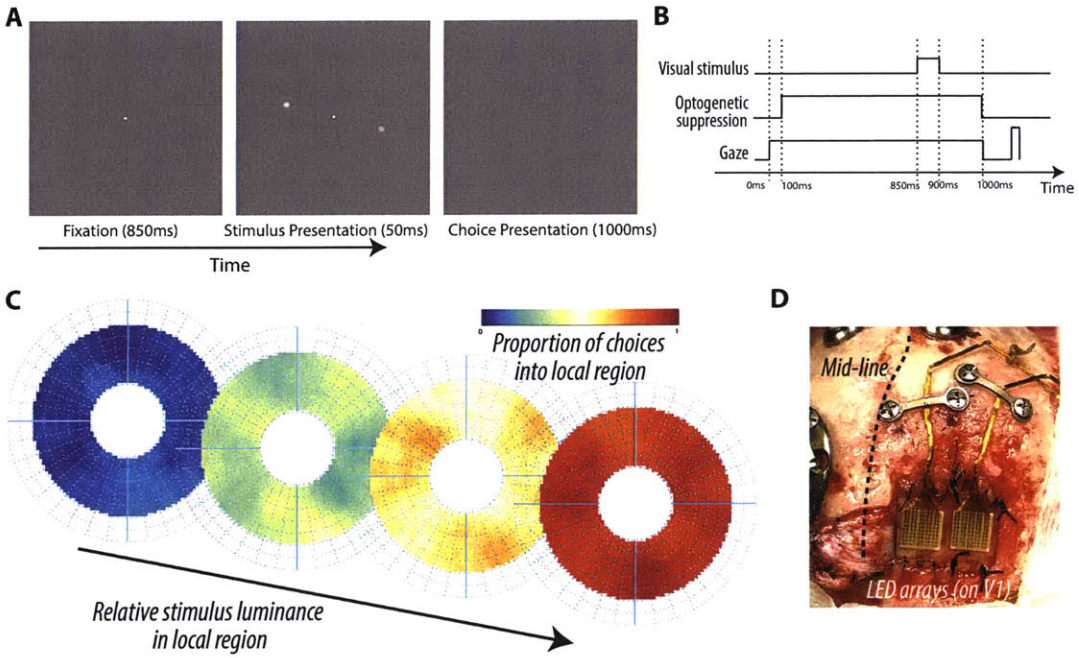

Figure 5-2: (a) Behavioral paradigm for luminance discrimination task. Each trial of the behavioral task consisted of a fixation period, during which one (or none) of the LEDs were preemptively activated on a random proportion of trials. Following fixation, two sample stimuli (Gaussian blob of 1° size, varying in luminance) were briefly presented at random radially opposite locations in the visually field. The task required the subject to make a saccade to a target location defined by the brighter of the two sample stimuli. The location and relative luminance of the stimuli was randomly assigned for each trial. By varying the relative luminance of the two sample stimuli, we systematically varied the task difficulty. (b) The time course of the behavioral paradigm. The LED activation was timed to completely overlap the stimulus-related activity in V1. (c) Each of the four discs correspond to the part of the peripheral visual field that was tested with this behavioral paradigm. The color of a given location (x,y) corresponds to the proportion of choices into a 1° pooling region centered at (x,y). Each panel corresponds to the relative stimulus luminance (also called signal) in a 1° pooling region centered at (x,y). As expected, the proportion of choices into a spatial region increases with increasing signal. (d) Photo of surgical implantation of two LED arrays over V1 cortex on the right hemishere.

Next we tested the efficacy of this tool for neural perturbation experiments that aim to constrain decoding models, i.e. perturbation for behavioral effects. We trained monkeys on a two-alternative-forced-choice (2AFC) luminance discrimination task (see Methods, Figure 5-2A,B). Briefly, each trial of the behavioral task consisted of a

central visual fixation period, followed by the simultaneous brief (50ms) presentation of two sample stimuli (Gaussian blob of 1 degree size, varying in luminance) in the periphery, at radially opposite locations in the visually field. The task required the subject to make a saccade to a target location defined by the brighter of the two sample stimuli. By varying the relative luminance of the two sample stimuli, we systematically varied the task difficulty. Stimuli were presented at randomly selected locations in the visual field within eccentricities of 3° to 10° of visual angle (resulting in a disc of tested visual space), only enforcing that the two stimuli were always radially opposite in position. As shown in Figure 5-2C, monkeys rapidly learned this task, and performed as expected (i.e. with increased choices into regions with increased signal).

We injected AAV8-CAG-ArchT on the right hemisphere of dorsal V1 in one monkey. Prior to implanting the LED arrays, we did not observe significant epifluorescence over the putatively transfected V1 cortex, suggesting poor viral expression. Thus, we first performed a small number of acute optrode experiments to verify viral expression and provide a baseline for comparison across methodologies. Figure 5-3 shows the behavioral effects in the two alternative forced choice luminance discrimination task described above, for an example optrode session. The colored disc corresponds to all tested regions of the visual space, where the center of the disc corresponds to the center of gaze, and each bin (x,y) corresponds to a location in the peripheral visual field where visual targets were presented; note that the foveal regions (eccentricity less than 3 degrees) were not tested. The color of each bin (x,y) indicates the change in psychophysical criterion when targets were near this location in the visual field (i.e. within a pooling zone of 3° diameter centered at this location), as a result of optogenetic suppression. Negative values indicate that the animal was biased away from this region (i.e. an increase in psychophysical threshold).

Given the anatomy of the perturbed cortical region (dorsal V1, right hemisphere), we expect behavioral effects to be spatially constrained to a target region of interest (target ROI) consisting of the contralateral lower visual field. Note that, given the symmetry of the task, this criterion change will also be present –with equal magnitude,

but opposite sign– in the radially opposite position in the visual field. The insets on the left show the psychometric curve fits for all trials sampled from the target ROI and a control ROI (contralateral upper visual field). Thus, the significant observed behavioral effects form this optrode experiment are consistent with the known causal role of V1 in the perception of luminance across the visual field. To localize putative effects, we optimized a Gaussian model with free parameters for the location, size and amplitude of a psychometric shift (see Methods). This model imposes a prior on spatially contiguous behavioral effects, consisting with the retinotopy of V1. The right-most panel of Figure 5-3 shows the Gaussian model fit for this psychometric shift map, with a localized effect in the contralateral lower visual field, and a negative but equal magnitude effect in the radially opposite region of visual space.

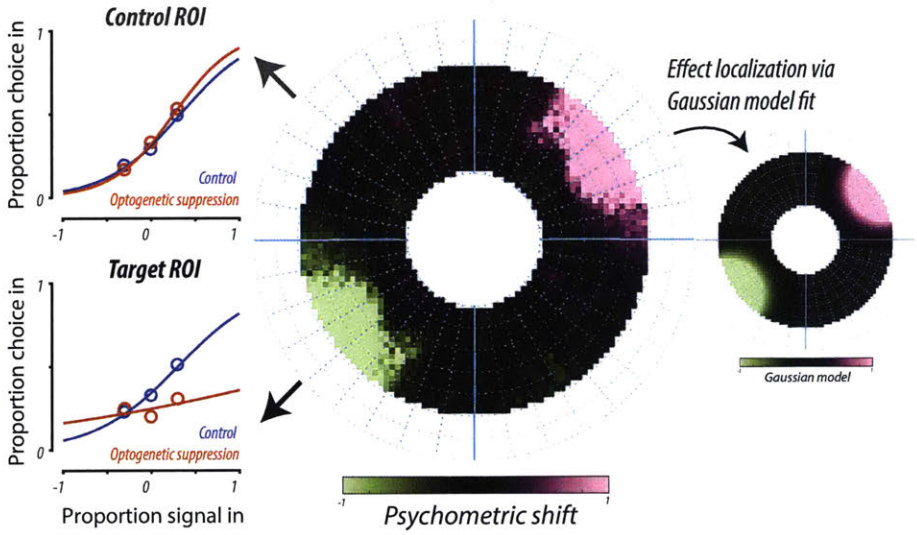

Figure 5-3: (a) Behavioral effects, corresponding to shifts of the subjects' psychometric curve, on luminance discrimination task from optogenetic suppression using acute optrodes (example session). Psychometric curves for the Behavioral effects were localized in a target ROI (contralateral lower visual field) by fitting a Gaussian model. For the

Next, we implanted two LED arrays over the transfected cortical tissue and repeated the same behavioral experiments, but with light delivery via the chronically implanted LEDs. For the first set of experiments, we activated groups of four neighbouring LEDs simultaneously to increase both the spatial spread and power of light.

We interleaved four such groups, each consisting of four corners of arrays. Given the chronic nature of this tool, we collected behavioral data over several sessions while activating LEDs on a small portion of trials (duty cycle = 20%). For an example activation condition of four neighbouring LEDs, Figure 5-4 shows the resulting psychometric shift maps; the inset shows the anatomical locations of each of the four activated LEDs, and the white circle overlaid on the effect map outlines to the localized effect, from fitting a Gaussian model. Plotting format is identical to Figure 5-3, except that pooling regions were significantly smaller in size (1°), given the larger number of trials. For this example LED condition, we observe a small putative shift in the psychometric curve, localized in the target ROI.

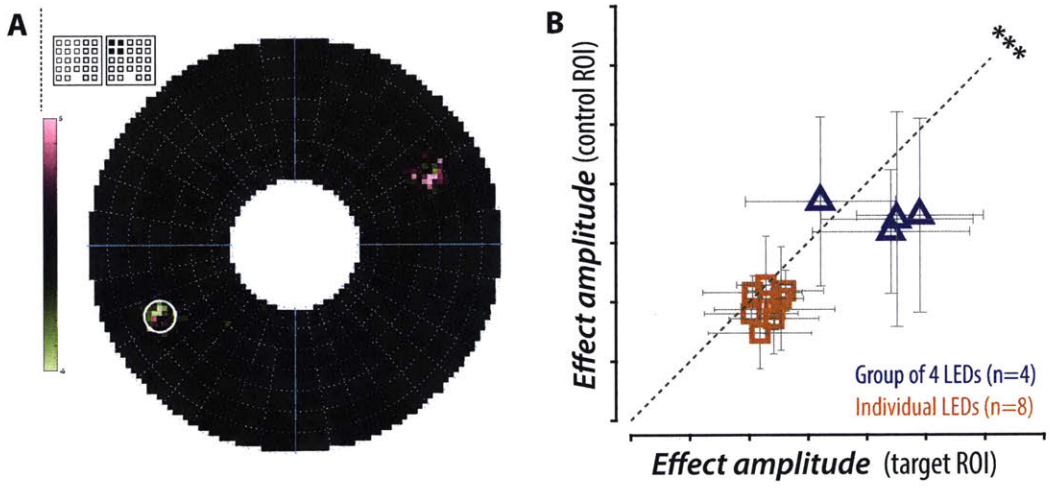

Figure 5-4: (a) Behavioral effects on luminance discrimination task from optogenetic suppression using chronically implanted LED array, for an example LED condition. The grid schematic (top right) shows which LEDs were activated for this condition. The white circle overlaid on the effect map corresponds to the localized effect, from fitting a Gaussian model. (b) Over all tested LED conditions, the amplitude of the localized effect, computed as the gain parameter of a Gaussian model fit, is significantly greater when localized effects are optimized within a target region of interest (ROI) as compared to a control region.

In total, we tested four such conditions (four simultaneously activated LEDs) and an additional eight conditions of activating individual LEDs. We collected these data over a large number of sessions, interleaving LED conditions within each of the two

subsets.

Over all tested LED conditions, the amplitude of the localized effect, computed as the gain parameter of a Gaussian model fit, is significantly greater when localized effects are optimized within a target ROI as compared to a control ROI ($p < 0.001$, exact test on distribution of median difference of effect amplitude between target and control ROI, estimated with bootstrap resampling over trials). From these data, we infer that neural suppression of V1 via activation of chronically implanted LEDs caused a localized change in the behavior of the animal and that this effect, over all tested LEDs, was reliable across trials. However, these data do not support inferences about the effects of activating individual LED activations, nor any dependences across LEDs (e.g. topographic organization of V1). We discuss these shortcomings in the Discussion, and refrain from making strong inferences about the success of this tool based on this data.

5.2.2 Perturbation of IT

Next we tested the efficacy of this tool to constrain decoding models of IT in core object recognition. We tested one monkey on a binary match-to-sample object recognition task (see Methods, Figure 5-5A). Briefly, each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image (6x6 degrees of visual angle in size) appeared at the center of gaze for 100ms. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right. One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms. Animals were rewarded with small juice rewards for successfully completing each trial. On each trial, one of ten binary object recognition tasks (listed in Figure 5-5A) was randomly selected.

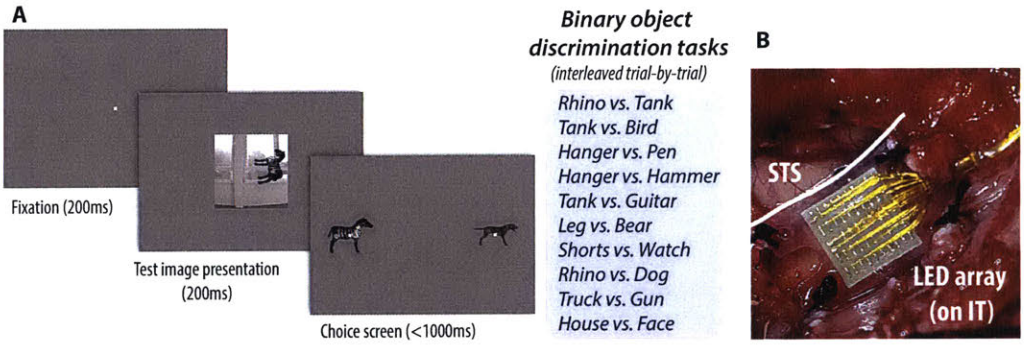

Figure 5-5: (a) Behavioral paradigm for object discrimination task. The list shows all ten tested pairwise object discrimination tasks, interleaved trial-by-trial. Each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image (6x6 degrees of visual angle in size) appeared at the center of gaze for 100ms. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right. One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms. Animals were rewarded with small juice rewards for successfully completing each trial. On each trial, one (or none) of the LEDs were preemptively activated on a random proportion of trials, timed to overlap with the feed-forward visual response in IT. (b) Photo of surgical implantation of one LED arrays over IT cortex on the left hemishere; STS corresponds to superior temporal sulcus.

We injected AAV8-CAG-ArchT on the lateral surface of IT on the left hemisphere of one monkey. After visualizing fluorescence, we implanted one LED array over this transfected tissue, as shown in Figure 5-5B. During experimental data collection, we first tested a large activation pattern consisting of every other LED (i.e. resembling a checkerboard), and activated this pattern on a random subset (30%) of trials for a total of seven sessions. Following this condition, we additionally tested seven individual LEDs, and activated one randomly assigned LED conditions on a random subset of trials (30% of trials were activation trials).

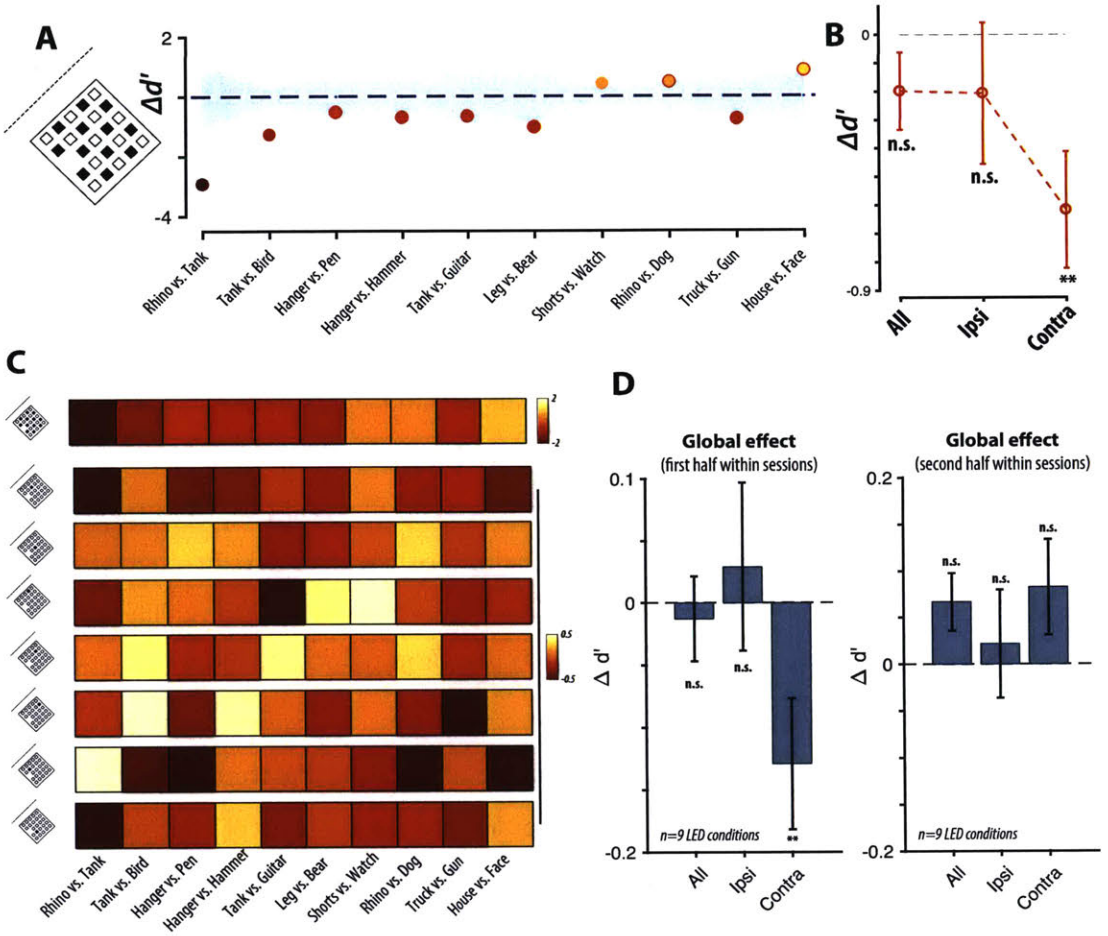

Figure 5-6: (a) Focusing on the first half of trials, the pattern of contralateral behavioral deficits (in units of d') over ten core object recognition tasks, for the checkerboard LED condition. The shaded region corresponds to the null distribution (obtained by randomly shuffling stimulation and control trials), while the colored dot corresponds to the observed deficit. (b) For the LED condition in (a), the global deficit (averaged over all tasks) for all, ipsilateral and contralateral stimuli. We report a significant global deficit for contralateral stimuli, but don't have the power to infer significant deficits on ipsilateral stimuli, or a difference in the deficit magnitude between ipsilateral and contralateral stimuli. (c) Patterns of deficits over ten core object recognition tasks, for each of the tested LED conditions. Each pattern is plotted as a heat maps where darker colors correspond to larger deficits, averaged over trials. The insets on the left of each heat map indicate which LED was activated. As in (a), these data correspond to the first half of trials for each behavioral session. (d) Corresponding to the deficit patterns shown in (c), the global deficit, averaged over all tasks and all LED conditions, is shown on the left panel. In contrast, the right panel shows the corresponding global deficit for the second half of trials for each behavioral session.

Over all trials, we observed no significant deficit in core object recognition performance resulting from neural suppression ($p > 0.05$, exact test on distribution of $\Delta d'$ averaged over all tasks and all tested LED conditions). This negative result could be due to many different potential factors, which are expanded upon in the Discussion. One such potential factor is the role of compensatory mechanisms on relatively fast time-scales (e.g. over tens of minutes, as observed in [Fetsch et al., 2018]). To investigate this possibility, we conducted post-hoc analyses by measuring behavioral deficits over trials in the first and second halves of trials within each behavioral session separately. Here, the first half of trials refers to pooled data from all behavioral sessions, including only the first half of trials from each session. Focusing on the first half of trials, we found significant task-selective deficits (see Figure 5-6A for contralateral behavioral deficits (in units of d') for the checkerboard LED condition). The shaded region corresponds to the null distribution (obtained by randomly shuffling stimulation and control trials), while the colored dot corresponds to the observe deficit. The inset on the left shows the pattern of activated LEDs. Contralateral stimuli refers to images where the center of the object was located on the side of the image that corresponds to the contralateral visual hemifield, while potentially still overlapping with both hemifields. Averaging over all tasks, there is a significant global contralateral deficit (see Figure 5-6B, $p < 0.01$). Note however that these data don't have the power to infer significant deficits on ipsilateral stimuli, or a difference in the deficit magnitude between ipsilateral and contralateral stimuli.

Figure 5-6C shows the patterns of deficits for each of the eight tested LED conditions as individual heat maps where darker colors correspond to larger deficits (mean over trials). These data do not have the power to support inferences about deficits over individual tasks and individual LEDs. However, the global deficit, averaged over all tasks and all LEDs, is shown in Figure 5-6D. Focusing on the first half of trials, we observe a significant global deficit for contralateral stimuli ($p < 0.01$, Figure 5-6, left panel). However, in the second half of trials within sessions, which pools all data from all behavioral sessions but including only the latter half of trials within each session, we observe no such deficit. These results are consistent the hypothesis of

transient behavioral effects which are attenuated over the course of the behavioral session. However, they could also reflect any number of sources of variability, including sampling variability from the null hypothesis, i.e. that LED activation led to no change in behavior. We expand on these possibilities in the Discussion.

5.3 Discussion

Motivated to increase the scale and throughput of current neural perturbation experiments in primates, we here tested a novel neural perturbation tool, a chronically implantable LED array for optogenetic perturbation. The data presented here do not support strong inferences about the utility of this tool, in its current form, for neural perturbation experiments in primates. Rather, these results provide a report of potentially promising preliminary findings with guides for improvements, both technological and experimental.

First, our photometric measurements suggest that this tool, when operated at a safe input intensity level, does not provide much light power to cortical neurons, and furthermore delivers light in a spatially diffuse manner. To the first point, computational models of light dispersion in cortical tissue [Chow et al., 2010] could provide insights into the necessary light power to match existing light delivery tools. Given that a major failure mode for this tool was electronic in nature (solder bond failure), simple electronic modifications could also drastically increase the safe operating input intensity. The coarse effective spatial resolution can likely be improved by focusing the light output of LEDs via miniaturized objectives. We speculate that the weak and unreliable putative behavioral effects observed in both sets of behavioral experiments stem primarily from this technical limit.

Specifically, in the first set of experiments (perturbing V1 while measuring luminance discrimination behavior), we observed behavioral effects consistent with an alternative method (acute optrodes), but with significant more variability. This "noise" may be due in part to additional behavioral variability (e.g. non-stationarity in the animal's behavior, captured by measuring behavior over several sessions) or even

include non-stationarity in the induced behavioral effects (e.g. due to long term compensation to the neural perturbation). In the second set of experiments (perturbing IT while measuring core object recognition behavior), we observed no significant behavioral deficits, in contrast to reliable, contralateral, relatively large deficits when neurons are suppressed with an alternative method (muscimol). Post-hoc analyses suggest the presence of transient behavioral deficits, which are attenuated over the course of each behavioral session. This phenomenology, similar to what was observed with optical stimulation of MT [Fetsch et al., 2018], is consistent with rapid downstream compensation. Importantly, we did not replicate this phenomenology in a second animal. Thus, this could also reflect any number of sources of variability, including sampling variability from the null hypothesis that LED activation led to no change in behavior.

However, we speculate that, in addition to the aforementioned technical limits, questions of causality may be fraught due to limitations of our current models of the neural phenomena under study, including aspects of plasticity, learning and compensation. For example, research in the domain of motion perception has shown that animal's training regime can directly alter the necessity of a brain region (MT) for a given task [Liu and Pack, 2017]. Moreover, such compensatory changes can even occur within a single behavioral session, and even within a single trial [Fetsch et al., 2018]. These phenomena are likely not specific to optogenetic perturbations, but the time-course of compensation may depend directly on the strength of the perturbation: weeks for lesions, days for potent pharmacological suppression, and minutes to hours for weaker perturbations (e.g. electrical, optogenetic). In addition to stronger perturbations, designing experiments that are robust to such changes— e.g. with very low proportion of perturbed trials, or with explicit behavioral washouts, or with perturbations aligned with "natural" neural activity—may be the key to obtain large-scale behavioral datasets with neural perturbation.

5.4 Methods

5.4.1 Subjects and surgery

Data presented were collected from two adult male rhesus macaque monkeys (Macaca mulatta, subjects Y, M). In both animals, a surgery using sterile technique was performed under general anaesthesia to implant a titanium head post to the skull using titanium screws.

Monkey Y was trained on a two-alternative forced-choice luminance discrimination task (see Figure 5-2). Following this, we injected AAV8-CAG-ArchT on the right hemisphere of primary visual (V1) cortex, covering a region of 15mmx7mm with over 18 injection sites. Over this transfected tissue, we first implanted a steel recording chamber (Crist) for acute optrode experiments, and confirmed weak viral expression by recording weak neural modulation by light. In a second surgery, we removed the chamber and implanted two 5x5 LED arrays over the transfected tissue. Arrays were held in place via dura sutures.

Monkey M was trained on a binary match-to-sample object discrimination task (see Figure 5-5), and previously implanted with a steel recording chamber (Crist) for acute electrophysiology experiments. We then injected AAV8-CAG-ArchT on the left hemisphere of IT cortex, covering a surface of 7mmx7mm with over 9 injection sites. We confirmed viral expression by visualizing epifluorescence. A single 5x5 LED array was implanted over the transfected tissue, and arrays were held in place via dura sutures.

All procedures were performed in compliance with National Institutes of Health guidelines and the standards of the MIT Committee on Animal Care and the American Physiological Society.

5.4.2 Behavioral paradigm and analysis

Luminance discrimination

The luminance discrimination behavioral task (see Figure 2B, 2C) was designed to probe the role of millimeter scale regions of V1 [Schiller and Tehovnik, 2003], which encode local features of the visual field. We trained two monkeys on this task, but here present data from one animal with sufficient number of trials per LED condition. Stimuli were presented on a 24" LCD monitor (1920 x 1080 at 60 Hz; Acer GD235HZ) and eye position was monitored by tracking the position of the pupil using a camera-based system (SR Research Eyelink 1000). At the start of each training session, the subject performed an eye-tracking calibration task by saccading to a range of spatial targets and maintaining fixation for 800 ms. Calibration was repeated if drift was noticed over the course of the session.

Each trial of the behavioral task consisted of a fixation period, during which one (or none) of the LEDs were preemptively activated on a random proportion of trials. Following fixation, two sample stimuli (Gaussian blob of 1° size, varying in luminance) were briefly presented at random radially opposite locations in the visually field. The LED activation was timed to completely overlap the stimulus-related activity in V1. The task required the subject to make a saccade to a target location defined by the brighter of the two sample stimuli. By varying the relative luminance of the two sample stimuli, we systematically varied the task difficulty.

To assess behavioral effects from stimulation, we first estimated psychometric curves from the animal's behavioral data, separately for each LED condition, and for each tested position in visual field. In other words, for each tested location (x, y), we pooled all trials within a 1° diameter zone centered at (x, y), and fitted a psychometric curve for each LED condition using logistic regression with two parameters: $\text{logit}(\text{choice}_{in}) = b_0 + b_1 \cdot \text{signal}_{in}$, where b_0, b_1 are the fitted parameters, and $\text{choice}_{in}, \text{signal}_{in}$ correspond to the dependent and experimentally controlled variables. For each psychometric curve, we defined the psychometric criterion as $-b_0/b_1$. To assess the effect of LED activation, we estimated the change in psychometric cri-

terion (i.e. corresponding to shifts in the psychometric curves) via the difference in estimated criterion between the corresponding curve fits. Repeating this procedure for each tested location in the visual field, we obtained a 2D map of psychometric shift estimates $\Psi(x, y)$. Rather than test each of these estimates independently, we fitted the entire 2D map $\Psi(x, y)$ via a Gaussian model:

$$\widetilde{\Psi}(x, y; A, \mu_x, \mu_y, \sigma) = A \times (e^{\frac{(x-\mu_x)^2+(y-\mu_y)^2}{\sigma^2}} - e^{\frac{(x+\mu_x)^2+(y+\mu_y)^2}{\sigma^2}}).$$

We constrained the variables μ_x, μ_y to reside inside a region of interest, thus imposing a prior on the effect of optogenetic suppression of V1. To infer whether LED activation led to any behavioral changes, we compared the distribution over trial-resampling of fitted amplitude parameters $|A|$ when the Gaussian model was constrained to localize an effect in the target ROI (contralateral lower visual field) versus a control ROI (contralateral upper visual field). To obtain a global statistic, we computed the median difference in model parameters over all LEDs.

Core object recognition

We examined basic-level, core object recognition behavior using a set of 25 broadly-sampled objects that we previously found to be reliably labeled by independent human subjects, based on the definition of basic-level proposed by [Rosch et al., 1976]. These images are identical to those used in Chapter 2; Figure 2-1 shows the list of 24 objects, with two example images of each object; we included one additional object (a face) for this set of experiments. From 300 possibly binary tasks, we tested a subsampled set of 10 tasks for these experiments. The images were sized so that they subtended $6 \times 6°$ for each monkey. Realtime experiments for all monkey psychophysics were controlled by open-source software (MWorks Project http://mworks-project.org/). Animals were rewarded with small juice rewards for successfully completing each trial, and received time-outs of 1.5 to 2.5 seconds for incorrect choices.

Monkey M was previously trained on a match-to-sample paradigm under head fixation and using gaze as the reporting effector. Eye position was monitored by

tracking the position of the pupil using a camera-based system (SR Research Eyelink II). Images were presented on a 27" LCD monitor (1920 x 1080 at 60 Hz; Samsung S27A850D) positioned 44 cm in front of the animal. At the start of each training session, subjects performed an eye-tracking calibration task by saccading to a range of spatial targets and maintaining fixation for 800ms. Calibration was repeated if drift was noticed over the course of the session.

Figure 5-5A illustrates the behavioral paradigm. Each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image appeared at the center of gaze for 100ms. Trials were aborted if gaze was not held within $\pm 2°$. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right (each centered at 6° of eccentricity along the horizontal meridian; see Fig. 1B). One of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms.

We computed behavioral deficits by measuring the difference in performance with respect to the one-versus-other object level performance metric (B.O2). Briefly, this metric is a pattern of pairwise object discrimination performances. For each pairwise object discrimination task, performance was estimated using a sensitivity index d' [Macmillan, 1993]: d' = Z(hit rate) - Z(false alarm rate), where Z(.) is the inverse of the cumulative Gaussian distribution. All d' estimates were constrained to a range of [0,5]. We additionally computed this deficit metric when restricting to contralateral/ipsilateral stimuli, i.e. images where the center of the object was located on the side of the image that corresponds to the contralateral/ipsilateral visual hemifield, while potentially still overlapping with both hemifields. Finally, to investigate the possibility of compensatory mechanisms on relatively fast time-scales (e.g. over tens of minutes, as observed in [Fetsch et al., 2018]), we also computed this metric when restricting to the first and second halves of trials within each behavioral session. Here,

the first half of trials refers to pooled data from all behavioral sessions, including only the first half of trials from each session.

Statistical testing

Unless otherwise specified, we estimated the uncertainty in delta measurements via bootstrap resampling of trials, repeated 100 times. The standard error of delta measurements was estimated as the standard deviation of this bootstrap distribution. For statistical tests, we performed one-tailed exact tests, by computing the empirical probability of observing a sample below zero. To compute this probability from the empirical bootstrap distribution, we fit a Gaussian kernel density function to the empirical distribution, optimizing the bandwidth parameter to minimize the mean squared error (kde.m on MATLAB file exchange). This kernel density function was evaluated to compute a p-value, by computing the cumulative probability of observing a positive behavioral delta.

5.5 Acknowledgements

This research was performed in collaboration with Arash Afraz and James J. DiCarlo. LED arrays were fabricated by BlackRock Microsystems.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Discussion

This thesis aims to provide quantitative insights into how the brain, in particular the highest level of ventral visual stream, causally supports basic-level core object recognition behavior. To date, a significant body of research suggests that IT cortex is a good candidate for the neural substrate of this behavior [Holmes and Gross, 1984, Horel et al., 1987, Biederman et al., 1997, Logothetis and Sheinberg, 1996, Tanaka, 1996, Rolls, 2000, DiCarlo et al., 2012, Kobatake and Tanaka, 1994, Ito et al., 1995, Logothetis et al., 1995, Booth and Rolls, 1998, Hung et al., 2005, Zhang et al., 2011, Sheinberg and Logothetis, 1997, de Beeck et al., 2001, Majaj et al., 2015]. Extending on this research program, we here sought to uncover if and how neuronal activity in IT *causally* supports basic-level core object recognition behavior. In addition to answering this qualitative question, we aim to provide new quantitative constraints for computational models of the ventral stream and core object recognition behavior. In Chapter 1, we defined quantitative understanding as uncovering computational models that quantitatively recapitulate all relevant phenomena, behavioral and neural, at a relevant level of abstract in the domain of core object recognition. Importantly, this process is iterative, as leading models strive to capture and synthesize all available and relevant observations, and make "predictions" within this domain of phenomena. In turn, new domains of phenomena (such as the ones presented in this work) enforce new constraints for future leading models. With this in mind, we here present the following advances towards this overarching goal.

6.1 Quantitative models of core object recognition

Our first subgoal can be stated as uncovering quantitative explanations for primate behavior in the domain of core object recognition. To this end, Chapters 2 and 3 established a scalable behavioral paradigm for testing the object recognition abilities of humans, monkeys and state-of-the-art models on hundreds of different object discrimination tasks. In Chapter 2, we systematically benchmarked the macaque monkey as a model of human visual processing in the domain of basic-level core object recognition. In Chapter 3, we applied these behavioral experiments to test the limits of state-of-the-art DCNNs. Using previously unattainable high-resolution behavioral metrics, we found that these models significantly diverged from primate behavior. These results suggest that new ANN models are needed to more precisely capture the neural mechanisms underlying primate object vision.

To this end, one strategy could be to use even larger-scale, high-resolution behavioral measurements, such as expanded versions of the patterns of image-level performance presented here, as useful top-down optimization guides. Not only do these high-resolution behavioral signatures have the statistical power to reject the currently leading ANN models, but they can also be efficiently collected at very large scale, via high-throughput psychophysical tools in humans and monkeys (e.g. Amazon Mechanical Turk for humans, Monkey Turk for rhesus monkeys), in contrast to other guide data (e.g. large-scale neuronal measurements, which are still under development). To be clear, we propose replacing or augmenting the model optimization procedure, which currently consists of optimizing for categorization performance on large scale image-sets, with optimization for a match between model and primate behavioral response patterns. As with current categorization-optimized models, claims of "generalization" to primate behavior and underlying neural responses in the ventral stream could be tested on held-out measurements, i.e. on new images or new images of new objects.

We intuit that models in a large space architectures can all approximate a given optimized target function, with differences in architectures corresponding to differ-

ences in sample efficiency. Thus, the success of such models, with respect to our neuroscience goals, may depend wholly on the similarity between the optimized target function and the neuroscientists' target function (e.g. ventral neuronal response patterns and behavioral patterns). The proposal can then be viewed as tightening the similarity between optimized target functions and neuroscience target functions by directly optimizing for neuroscience target functions. This strategy is complementary to approaches involving sampling over model architectures with a fixed optimization procedure [Zoph et al., 2017, Yamins et al., 2014, Rajalingham et al., 2018, Jozwik et al., 2016, Kheradpisheh et al., 2016, Kubilius et al., 2016, Cadieu et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014]. Based on our current–albeit preliminary– results, we speculate that this strategy is less likely to quickly yield primate-like models, as compared to sampling over optimization procedures. This speculation is grounded by the intuition stated above, and supported in part by observations that the precise set of images and labels used for model training significantly impact the resulting model features and corresponding behavioral responses (see Figure 3-4, where training a fixed model architecture on different choices of image sets leads to vastly different pattern of residuals relative to humans). In contrast, choices of architectures within this model class have little effect (see Figure 3-4). An important caveat is that comparisons between variations in model architecture and variations in model training data cannot be matched ("apples to apples'), and furthermore that models sampled from a larger space of architectures may show greater differences.

6.2 The causal role of IT cortex in core object recognition

Our second subgoal can be stated as obtaining direct causal evidence for the role of IT in core object recognition behavior. As discussed in Chapter 1, inferring causal dependencies between different phenomena relies on experimental control of one these phenomena. More precisely, the confidence of inferred causal dependencies between

measured phenomena X', Y depends on the (estimated) equivalence between the experimentally controlled variable X and the inferred causal variable X'. In Chapter 4, we used a well established neural suppression agent (muscimol) to reversibly inactivate individual, arbitrarily sampled millimeter-scale regions of IT while monkeys performed a battery of binary core object discrimination tasks, interleaved trial-by-trial. Our results results provide much needed direct causal evidence for the general decoding hypothesis and, importantly, are the first to demonstrate the necessity of IT cortex for a wide range of general core object recognition behaviors with behaviorally critical topographic organization.

Moreover, our results provide rich constraints for computational models of the neural mechanisms underlying primate behavior in the domain of core object recognition. To date, leading ANN models of the ventral stream have been tested largely on the available quantitative phenomena, consisting of neural and behavioral responses to images. Importantly, our goal is not simply to produce a model that captures image responses at various levels, but rather to produce a model that captures all relevant phenomena. As such, perturbations with quantitative measures of the resulting neural and behavioral responses are a rich new domain of constraints for computational models. In this work, we first measured the impact of millimeter-scale perturbations on a battery of core object recognition tasks, both in terms of global magnitude of deficit and finer-grained characteristics, such as the magnitude of deficits over various subsets of images (see Figure 4-3, e.g. contralateral stimuli) and sparsity of deficits over tasks (see Figure 4-4). Second, we observed that effects of inactivation were topographically organized and measured the spatial auto-correlation of the read-out of IT for behavior. Finally, we quantitatively compared several read-out models that map the neuronal activity patterns at local regions to predicted inactivation effects on behavior, and found that behavioral deficits are predicted by the local neuronal selectivity, rather than response — a new constraint for the mapping from neurons to behavior.

Importantly, these results are powerful constraints for constructing new ANN models of the primate ventral visual stream and core object recognition behavior.

Given that current ANN models do not capture the vast majority of these phenomena (not shown), this motivates constructing new ANNs with precise mapping of computational layers and features to anatomical mechanisms in the primate brain, which we refer to as "topographic deep artificial neural networks" (TDANNs). Preliminary work suggests that a first generation of TDANNs recapitulate some but not all first-order experimental phenomena better than current deep ANN models. This research establishes a new class of experimental constraints for computational models of core object recognition.

6.3 Future goals

In Chapter 5, we tested a novel chronically implantable array of LEDs for optogenetic experiments in primates. Chronic tools such as this one are promising avenues for high-throughput behavioral experiments with time-delimited perturbation of neural activity, which could enable the collection of a new class of large-scale behavior datasets and corresponding powerful constraints for models of object recognition. While this particular tool is far from high-throughput in its current implementation, there are clear possible technological improvements that may increase its utility for our systems neuroscience goals.

As observed in Chapter 5, questions of causality may be fraught due to limitations of our current models of the phenomena under study, including aspects of plasticity, learning and compensation. For example, research in the domain of motion perception has shown that animal's training regime can directly alter the necessity of a brain region (MT) for a given task [Liu and Pack, 2017]. Moreover, such compensatory changes can even occur within a single behavioral session, and even within a single trial [Fetsch et al., 2018]. In the domain of visual object recognition, direct manipulations of IT in general visual recognition behavior have also largely been equivocal. Lesions of IT sometimes suggest the necessity of IT and visual behaviors [Cowey and Gross, 1970, Manning, 1972, Holmes and Gross, 1984, Biederman et al., 1997, Buffalo et al., 2000] but the resulting behavioral deficits are often contradictory (with often

no lasting visual deficits) [Dean, 1974, Huxlin et al., 2000] and at best modest (e.g. 10-15% drop in performance for large-scale bilateral removal of IT when a complete loss of performance would have been 40%) [Horel et al., 1987, Matsumoto et al., 2016]. This is in stark contrast to relatively large effects for relatively small ($\sim$ 2mm diameter) perturbations of IT [Afraz et al., 2006, Afraz et al., 2015, Moeller et al., 2017, Sadagopan et al., 2017], as well to the current work. Additionally, we previously observed that a medium-scale ($\sim$ 9mm diameter) inactivation of face-selective regions in IT resulted in complete deficit for contralateral face discrimination behavior (not shown). What explains these discrepancies across studies?

It is thought that this variability could be due to differences in the behavioral tasks that were tested, and corresponding available alternative strategies [DiCarlo et al., 2012]. To this end, a key future goal would be to formalize and test this intuition by constructing a single model, including not only the topographic mapping of features but also the mechanisms underlying compensation and learning across a wide range of behavioral tasks, that synthesizes all available observations from perturbations of IT. Concretely, such a model would not only predict the magnitude/sparsity/etc. of behavioral deficits as a function of the inactivation size, but also reflect the dynamics of compensatory mechanisms, which can subtend from milliseconds to tens of trials to weeks. While this may seem like an ambitious goal, it is possible that modifications to current models, e.g. topographic ANN models with reinforcement learning, capture a large portion of the observed phenomena.

Bibliography

[Afraz et al., 2015] Afraz, A., Boyden, E. S., and DiCarlo, J. J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences*, 112(21):6730–6735.

[Afraz et al., 2006] Afraz, S.-R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442(7103):692–695.

[Andrews and Johnston, 1979] Andrews, P. and Johnston, G. (1979). Gaba agonists and antagonists. *Biochemical pharmacology*, 28(18):2697–2702.

[Arikan et al., 2002] Arikan, R., Blake, N. M., Erinjeri, J. P., Woolsey, T. A., Giraud, L., and Highstein, S. M. (2002). A method to measure the effective spread of focally injected muscimol into the central nervous system with electrophysiology and light microscopy. *Journal of neuroscience methods*, 118(1):51–57.

[Biederman et al., 1997] Biederman, I., Gerhardstein, P. C., Cooper, E. E., and Nelson, C. A. (1997). High level object recognition without an anterior inferior temporal lobe. *Neuropsychologia*, 35(3):271–287.

[Bisley et al., 2001] Bisley, J. W., Zaksas, D., and Pasternak, T. (2001). Microstimulation of cortical area mt affects performance on a visual working memory task. *Journal of neurophysiology*, 85(1):187–196.

[Booth and Rolls, 1998] Booth, M. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8(6):510–523.

[Britten and van Wezel, 1998] Britten, K. H. and van Wezel, R. J. (1998). Electrical microstimulation of cortical area mst biases heading perception in monkeys. *Nature neuroscience*, 1(1):59–63.

[Brown, 1910] Brown, W. (1910). Some experimental results in the correlation of mental abilities1. *British Journal of Psychology, 1904-1920*, 3(3):296–322.

[Bruce, 1982] Bruce, C. (1982). Face recognition by monkeys: absence of an inversion effect. *Neuropsychologia*, 20(5):515–521.

[Bruce et al., 1981] Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol*, 46(2):369–384.

[Buffalo et al., 2000] Buffalo, E. A., Ramus, S. J., Squire, L. R., and Zola, S. M. (2000). Perception and recognition memory in monkeys following lesions of area te and perirhinal cortex. *Learning & Memory*, 7(6):375–382.

[Buffalo et al., 1998] Buffalo, E. A., Stefanacci, L., Squire, L. R., and Zola, S. M. (1998). A reexamination of the concurrent discrimination learning task: the importance of anterior inferotemporal cortex, area te. *Behavioral neuroscience*, 112(1):3.

[Cadena et al., 2017] Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2017). Deep convolutional models improve predictions of macaque v1 responses to natural images. *bioRxiv*, page 201764.

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

[Cavanaugh et al., 2012] Cavanaugh, J., Monosov, I. E., McAlonan, K., Berman, R., Smith, M. K., Cao, V., Wang, K. H., Boyden, E. S., and Wurtz, R. H. (2012). Optogenetic inactivation modifies monkey visuomotor behavior. *Neuron*, 76(5):901–907.

[Celebrini and Newsome, 1995] Celebrini, S. and Newsome, W. T. (1995). Microstimulation of extrastriate area mst influences performance on a direction discrimination task. *Journal of Neurophysiology*, 73(2):437–448.

[Chang and Tsao, 2017] Chang, L. and Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028.

[Chklovskii et al., 2002] Chklovskii, D. B., Schikorski, T., and Stevens, C. F. (2002). Wiring optimization in cortical circuits. *Neuron*, 34(3):341–347.

[Chow et al., 2010] Chow, B. Y., Han, X., Dobry, A. S., Qian, X., Chuong, A. S., Li, M., Henninger, M. A., Belfort, G. M., Lin, Y., Monahan, P. E., et al. (2010). High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature*, 463(7277):98.

[Cichy et al., 2016] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755.

[Conway et al., 2007] Conway, B. R., Moeller, S., and Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron*, 56(3):560–573.

[Cowey and Gross, 1970] Cowey, A. and Gross, C. (1970). Effects of foveal prestriate and inferotemporal lesions on visual discrimination by rhesus monkeys. *Experimental Brain Research*, 11(2):128–144.

[Cox et al., 2008] Cox, D. D., Papanastassiou, A. M., Oreper, D., Andken, B. B., and DiCarlo, J. J. (2008). High-resolution three-dimensional microelectrode brain mapping using stereo microfocal x-ray imaging. *Journal of Neurophysiology*, 100(5):2966–2976.

[Crump et al., 2013] Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410.

[Dai et al., 2014] Dai, J., Brooks, D. I., and Sheinberg, D. L. (2014). Optogenetic and electrical microstimulation systematically bias visuospatial choice in primates. *Current Biology*, 24(1):63–69.

[de Beeck et al., 2001] de Beeck, H. O., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature neuroscience*, 4(12):1244.

[De Valois et al., 1974a] De Valois, R. L., Morgan, H., and Snodderly, D. M. (1974a). Psychophysical studies of monkey vision-iii. spatial luminance contrast sensitivity tests of macaque and human observers. *Vision research*, 14(1):75–81.

[De Valois et al., 1974b] De Valois, R. L., Morgan, H. C., Polson, M. C., Mead, W. R., and Hull, E. M. (1974b). Psychophysical studies of monkey visionâĂŤi. macaque luminosity and color vision tests. *Vision research*, 14(1):53–67.

[Dean, 1974] Dean, P. (1974). The effect of inferotemporal lesions on memory for visual stimuli in rhesus monkeys. *Brain research*, 77(3):451–469.

[DeAngelis et al., 1998] DeAngelis, G. C., Cumming, B. G., and Newsome, W. T. (1998). Cortical area mt and the perception of stereoscopic depth. *Nature*, 394(6694):677.

[Deisseroth, 2011] Deisseroth, K. (2011). Optogenetics. *Nature methods*, 8(1):26–29.

[DiCarlo and Cox, 2007] DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341.

[DiCarlo and Johnson, 1999] DiCarlo, J. J. and Johnson, K. O. (1999). Velocity invariance of receptive field structure in somatosensory cortical area 3b of the alert monkey. *Journal of Neuroscience*, 19(1):401–419.

[DiCarlo et al., 2012] DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

[Dodge and Karam, 2017] Dodge, S. and Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. *arXiv preprint arXiv:1705.02498.*

[Eldridge et al., 2016] Eldridge, M. A., Lerchner, W., Saunders, R. C., Kaneko, H., Krausz, K. W., Gonzalez, F. J., Ji, B., Higuchi, M., Minamimoto, T., and Richmond, B. J. (2016). Chemogenetic disconnection of monkey orbitofrontal and rhinal cortex reversibly disrupts reward value. *Nature neuroscience*, 19(1):37.

[Felleman and Van Essen, 1991] Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47.

[Fetsch et al., 2018] Fetsch, C. R., Odean, N. N., Jeurissen, D., El-Shamayleh, Y., Horwitz, G. D., and Shadlen, M. N. (2018). Focal optogenetic suppression in macaque area mt biases direction discrimination and choice confidence, but only transiently. *bioRxiv*, page 277251.

[Fujita et al., 1992] Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360(6402):343–346.

[Gagin et al., 2014] Gagin, G., Bohon, K. S., Butensky, A., Gates, M. A., Hu, J.-Y., Lafer-Sousa, R., Pulumo, R. L., Qu, J., Stoughton, C. M., and Swanbeck, S. N. (2014). Color-detection thresholds in rhesus macaque monkeys and humans. *Journal of vision*, 14(8):12.

[Geirhos et al., 2017] Geirhos, R., Janssen, D. H., SchÃijtt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969.*

[Gerits et al., 2012] Gerits, A., Farivar, R., Rosen, B. R., Wald, L. L., Boyden, E. S., and Vanduffel, W. (2012). Optogenetically induced behavioral and functional network changes in primates. *Current Biology*, 22(18):1722–1726.

[Gibson, 1979] Gibson, J. J. (1979). *The ecological approach to visual perception: classic edition.* Psychology Press.

[Goodale and Milner, 1992] Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25.

[Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572.*

[Grill-Spector and Weiner, 2014] Grill-Spector, K. and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548.

[Gross, 1994] Gross, C. G. (1994). How inferior temporal cortex became a visual area. *Cerebral cortex*, 4(5):455–469.

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

[Han et al., 2011] Han, X., Chow, B. Y., Zhou, H., Klapoetke, N. C., Chuong, A., Rajimehr, R., Yang, A., Baratta, M. V., Winkle, J., and Desimone, R. (2011). A high-light sensitivity optical neural silencer: development and application to optogenetic control of non-human primate cortex. *Frontiers in systems neuroscience*, 5.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Hegdé and Van Essen, 2000] Hegdé, J. and Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area v2. *Journal of Neuroscience*, 20(5):RC61–RC61.

[Holmes and Gross, 1984] Holmes, E. J. and Gross, C. G. (1984). Effects of inferior temporal lesions on discrimination of stimuli differing in orientation. *The Journal of Neuroscience*, 4(12):3063–3068.

[Hong et al., 2016] Hong, H., Yamins, D. L., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622.

[Horel et al., 1987] Horel, J. A., Pytko-Joiner, D. E., Voytko, M. L., and Salsbury, K. (1987). The performance of visual tasks while segments of the inferotemporal cortex are suppressed by cold. *Behavioural brain research*, 23(1):29–42.

[Hosseini et al., 2017] Hosseini, H., Xiao, B., Jaiswal, M., and Poovendran, R. (2017). On the limitation of convolutional neural networks in recognizing negative images. *human performance*, 4(5):6.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154.

[Hubel and Wiesel, 1968] Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

[Hung et al., 2005] Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866.

[Huxlin et al., 2000] Huxlin, K. R., Saunders, R. C., Marchionini, D., Pham, H.-A., and Merigan, W. H. (2000). Perceptual deficits after lesions of inferotemporal cortex in macaques. *Cerebral Cortex*, 10(7):671–683.

[Huynh, 2009] Huynh, D. Q. (2009). Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164.

[Issa et al., 2013] Issa, E. B., Papanastassiou, A. M., and DiCarlo, J. J. (2013). Large-scale, high-resolution neurophysiological maps underlying fmri of macaque temporal lobe. *Journal of Neuroscience*, 33(38):15207–15219.

[Ito et al., 1995] Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of neurophysiology*, 73(1):218–226.

[Jazayeri and Afraz, 2017] Jazayeri, M. and Afraz, A. (2017). Navigating the neural space in search of the neural code. *Neuron*, 93(5):1003–1014.

[Jazayeri et al., 2012] Jazayeri, M., Lindbloom-Brown, Z., and Horwitz, G. D. (2012). Saccadic eye movements evoked by optogenetic activation of primate v1. *Nature neuroscience*, 15(10):1368–1370.

[Johnson et al., 2002] Johnson, K. O., Hsiao, S. S., and Yoshioka, T. (2002). Neural coding and the basic law of psychophysics. *The Neuroscientist*, 8(2):111–121.

[Jozwik et al., 2016] Jozwik, K. M., Kriegeskorte, N., and Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. *Neuropsychologia*, 83:201–226.

[Katz et al., 2016] Katz, L. N., Yates, J. L., Pillow, J. W., and Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, 535(7611):285.

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

[Kheradpisheh et al., 2016] Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6:32672.

[Kiorpes et al., 2008] Kiorpes, L., Li, D., and Hagan, M. (2008). Crowding in primates: a comparison of humans and macaque monkeys. *Perception*, 37:37.

[Kobatake and Tanaka, 1994] Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3):856–867.

[Kornblith et al., 2013] Kornblith, S., Cheng, X., Ohayon, S., and Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron*, 79(4):766–781.

[Kravitz et al., 2011] Kravitz, D. J., Saleem, K. S., Baker, C. I., and Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217.

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

[Kriegeskorte et al., 2009] Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[Kubilius et al., 2016] Kubilius, J., Bracci, S., and de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896.

[Kumar and Hedges, 1998] Kumar, S. and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution. *Nature*, 392(6679):917.

[Lafer-Sousa and Conway, 2013] Lafer-Sousa, R. and Conway, B. R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature neuroscience*, 16(12):1870.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[Liu and Pack, 2017] Liu, L. D. and Pack, C. C. (2017). The contribution of area mt to visual motion perception depends on training. *Neuron*, 95(2):436–446.

[Logothetis et al., 1995] Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.

[Logothetis and Sheinberg, 1996] Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annual review of neuroscience*, 19(1):577–621.

[Macmillan, 1993] Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model.

[Majaj et al., 2015] Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *The Journal of Neuroscience*, 35(39):13402–13418.

[Manning, 1972] Manning, F. J. (1972). Serial reversal learning by monkeys with inferotemporal or foveal prestriate lesions. *Physiology and behavior*, 8(2):177–181.

[Mantini et al., 2012] Mantini, D., Hasson, U., Betti, V., Perrucci, M. G., Romani, G. L., Corbetta, M., Orban, G. A., and Vanduffel, W. (2012). Interspecies activity correlations reveal functional correspondence between monkey and human brain areas. *Nature methods*, 9(3):277–282.

[Matsumoto et al., 2016] Matsumoto, N., Eldridge, M. A., Saunders, R. C., Reoli, R., and Richmond, B. J. (2016). Mild perceptual categorization deficits follow bilateral removal of anterior inferior temporal cortex in rhesus monkeys. *Journal of Neuroscience*, 36(1):43–53.

[Minamimoto et al., 2010] Minamimoto, T., Saunders, R. C., and Richmond, B. J. (2010). Monkeys quickly learn and generalize visual categories without lateral prefrontal cortex. *Neuron*, 66(4):501–507.

[Miranda-Dominguez et al., 2014] Miranda-Dominguez, O., Mills, B. D., Grayson, D., Woodall, A., Grant, K. A., Kroenke, C. D., and Fair, D. A. (2014). Bridging the gap between the human and macaque connectome: a quantitative comparison of global interspecies structure-function relationships and network topology. *The Journal of Neuroscience*, 34(16):5552–5563.

[Mishkin et al., 1983] Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

[Miyashita, 1993] Miyashita, Y. (1993). Inferior temporal cortex: where visual perception meets memory. *Annual review of neuroscience*, 16(1):245–263.

[Moeller et al., 2017] Moeller, S., Crapse, T., Chang, L., and Tsao, D. Y. (2017). The effect of face patch microstimulation on perception of faces and objects. *Nature Neuroscience*, 20(5):743–752.

[Mutch and Lowe, 2008] Mutch, J. and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57.

[Nguyen et al., 2015] Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436.

[Nichols and Newsome, 2002] Nichols, M. J. and Newsome, W. T. (2002). Middle temporal visual area microstimulation influences veridical judgments of motion direction. *Journal of Neuroscience*, 22(21):9530–9540.

[Noudoost and Moore, 2011] Noudoost, B. and Moore, T. (2011). A reliable microinjectrode system for use in behaving monkeys. *Journal of neuroscience methods*, 194(2):218–223.

[Ohayon et al., 2013] Ohayon, S., Grimaldi, P., Schweers, N., and Tsao, D. Y. (2013). Saccade modulation by optical and electrical stimulation in the macaque frontal eye field. *The Journal of Neuroscience*, 33(42):16684–16697.

[Okamura et al., 2014] Okamura, J.-y., Yamaguchi, R., Honda, K., Wang, G., and Tanaka, K. (2014). Neural substrates of view-invariant object recognition developed without experiencing rotations of the objects. *The Journal of Neuroscience*, 34(45):15047–15059.

[Op De Beeck and Vogels, 2000] Op De Beeck, H. and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology*, 426(4):505–518.

[Orban et al., 2004] Orban, G. A., Van Essen, D., and Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in cognitive sciences*, 8(7):315–324.

[Papert, 1966] Papert, S. A. (1966). The summer vision project.

[Parr, 2011] Parr, L. A. (2011). The inversion effect reveals species differences in face processing. *Acta psychologica*, 138(1):204–210.

[Passingham, 2009] Passingham, R. (2009). How good is the macaque monkey model of the human brain? *Current opinion in neurobiology*, 19(1):6–11.

[Pasupathy and Connor, 2002] Pasupathy, A. and Connor, C. E. (2002). Population coding of shape in area v4. *Nature neuroscience*, 5(12):1332.

[Peterson et al., 2016] Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*.

[Pinto et al., 2008] Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27.

[Rajalingham and DiCarlo, 2018] Rajalingham, R. and DiCarlo, J. (2018). Reversible inactivation of different millimeter-scale regions of primate it results in different patterns of core object recognition deficits. in preparation.

[Rajalingham et al., 2018] Rajalingham, R., Issa, E., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. biorxiv, 240614.

[Rajalingham et al., 2015] Rajalingham, R., Schmidt, K., and DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *The Journal of Neuroscience*, 35(35):12127–12136.

[Recanzone et al., 1992] Recanzone, G., Merzenich, M., and Dinse, H. (1992). Expansion of the cortical representation of a specific skin field in primary somatosensory cortex by intracortical microstimulation. *Cerebral Cortex*, 2(3):181–196.

[RichardWebster et al., 2016] RichardWebster, B., Anthony, S. E., and Scheirer, W. J. (2016). Psyphy: A psychophysics driven evaluation framework for visual recognition. *arXiv preprint arXiv:1611.06448*.

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

[Rolls, 2000] Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–218.

[Romo et al., 2000] Romo, R., Hernandez, A., Zainos, A., Brody, C. D., and Lemus, L. (2000). Sensing without touching: psychophysical performance based on cortical microstimulation. *Neuron*, 26(1):273–278.

[Romo et al., 1998] Romo, R., Hernandez, A., Zainos, A., and Salinas, E. (1998). Somatosensory discrimination based on cortical microstimulation. *Nature*, 392(6674):387–390.

[Rosch et al., 1976] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

[Rust and DiCarlo, 2010] Rust, N. C. and DiCarlo, J. J. (2010). Selectivity and tolerance (âĂIJinvarianceâĂİ) both increase as visual information propagates from cortical area v4 to it. *The Journal of Neuroscience*, 30(39):12978–12995.

[Sadagopan et al., 2017] Sadagopan, S., Zarco, W., and Freiwald, W. A. (2017). A causal relationship between face-patch activity and face-detection behavior. *eLife*, 6:e18558.

[Salzman et al., 1990] Salzman, C. D., Britten, K. H., and Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346(6280):174–177.

[Schiller and Tehovnik, 2003] Schiller, P. H. and Tehovnik, E. J. (2003). Cortical inhibitory circuits in eye-movement generation. *European Journal of Neuroscience*, 18(11):3127–3133.

[Seibert et al., 2016] Seibert, D., Yamins, D. L., Ardila, D., Hong, H., DiCarlo, J. J., and Gardner, J. L. (2016). A performance-optimized model of neural responses across the ventral visual stream. *bioRxiv*, page 036475.

[Serre et al., 2007] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426.

[Sheinberg and Logothetis, 1997] Sheinberg, D. L. and Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences*, 94(7):3408–3413.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Spearman, 1910] Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295.

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

[Tanaka, 1996] Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139.

[Thier and Andersen, 1998] Thier, P. and Andersen, R. A. (1998). Electrical microstimulation distinguishes distinct saccade-related areas in the posterior parietal cortex. *Journal of Neurophysiology*, 80(4):1713–1735.

[Tsao et al., 2003] Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nature neuroscience*, 6(9):989–995.

[Tsao et al., 2006] Tsao, D. Y., Freiwald, W. A., Tootell, R. B., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674.

[Tsao and Livingstone, 2008] Tsao, D. Y. and Livingstone, M. S. (2008). Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31:411–437.

[Tsunoda et al., 2001] Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature neuroscience*, 4(8):832–838.

[Ullman, 2000] Ullman, S. (2000). *High-level vision: Object recognition and visual cognition.* MIT press.

[Ullman and Humphreys, 1996] Ullman, S. and Humphreys, G. W. (1996). *High-level vision: Object recognition and visual cognition*, volume 2. MIT press Cambridge, MA.

[Van Essen, 1979] Van Essen, D. C. (1979). Visual areas of the mammalian cerebral cortex. *Annual Review of Neuroscience*, 2(1):227–261.

[Vazquez et al., 2000] Vazquez, P., Cano, M., and AcuÃśa, C. (2000). Discrimination of line orientation in humans and monkeys. *Journal of Neurophysiology*, 83(5):2639–2648.

[Verhoef et al., 2015] Verhoef, B.-E., Bohon, K. S., and Conway, B. R. (2015). Functional architecture for disparity in macaque inferior temporal cortex and its relationship to the architecture for faces, color, scenes, and visual field. *Journal of Neuroscience*, 35(17):6952–6968.

[Verhoef et al., 2012] Verhoef, B.-E., Vogels, R., and Janssen, P. (2012). Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron*, 73(1):171–182.

[Vermeire and Hamilton, 1998] Vermeire, B. A. and Hamilton, C. R. (1998). Inversion effect for faces in split-brain monkeys. *Neuropsychologia*, 36(10):1003–1014.

[Vinje and Gallant, 2000] Vinje, W. E. and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276.

[Vogels and Orban, 1990] Vogels, R. and Orban, G. (1990). How well do response changes of striate neurons signal differences in orientation: a study in the discriminating monkey. *The Journal of neuroscience*, 10(11):3543–3558.

[Wallis et al., 2017] Wallis, T. S., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., and Bethge, M. (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of vision*, 17(12):5–5.

[Wan et al., 2013] Wan, L., Zeiler, M., Zhang, S., Cun, Y. L., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1058–1066.

[Wang et al., 1996] Wang, G., Tanaka, K., and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, 272(5268):1665.

[Wang et al., 1998] Wang, G., Tanifuji, M., and Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience research*, 32(1):33–46.

[Weiskrantz and Saunders, 1984] Weiskrantz, L. and Saunders, R. (1984). Impairments of visual object transforms in monkeys. *Brain*, 107(4):1033–1072.

[Wen et al., 2017] Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, pages 1–25.

[Yamins and DiCarlo, 2016] Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

[Yamins et al., 2014] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, page 201403112.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). *Visualizing and understanding convolutional networks*, pages 818–833. Springer.

[Zhang et al., 2011] Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences*, 108(21):8850–8855.

[Zoccolan et al., 2009] Zoccolan, D., Oertelt, N., DiCarlo, J. J., and Cox, D. D. (2009). A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences*, 106(21):8748–8753.

[Zoph et al., 2017] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2017). Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*.