bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks

Abbreviated title: Comparing object recognition behavior in humans, monkeys, and machines

Rishi Rajalingham*, Elias B. Issa*, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo

McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

*R.R. and E.B.I. contributed equally to this work.

Correspondence should be addressed to James J. DiCarlo, McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Institute of Technology, 46-6161, Cambridge, MA 02139. E-mail: dicarlo@mit.edu

E. Issa's present address: Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027

Targeting *Journal of Neuroscience*
Title 19 words
Abbreviated Title 57 characters
Abstract 291 words
Significance Statement 97 words
Introduction 895 words
Discussion 1051words
Figures 6
*All word limits include citations

AUTHOR CONTRIBUTIONS

E.B.I., R.R., and J.J.D designed the experiments. E.B.I., K.S., R.R., and K.K. carried out the experiments. R.R., E.B.I., and P.B. performed the data analysis and modeling. R.R., E.B.I., and J.J.D. wrote the manuscript.

ACKNOWLEDGEMENTS

This research was supported by US National Eye Institute grants R01-EY014970 (J.J.D.) and K99-EY022671 (E.B.I.), Office of Naval Research MURI-114407 (J.J.D), Engineering Research Council of Canada (R.R.), and The McGovern Institute for Brain Research

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Primates—including humans—can typically recognize objects in visual images at a glance, even in the face of naturally occurring identity preserving image transformations such as changes in viewpoint. A primary neuroscience goal is to uncover neuron-level mechanistic models that quantitatively explain this behavior, not only predicting average primate performance, but also predicting primate performance for each and every image. Here, we applied this stringent behavioral prediction test to the leading mechanistic models of primate vision (specifically, deep, convolutional, artificial neural networks; ANNs) by directly comparing their behavioral patterns, at high resolution over a large number of object discrimination tasks, against those of humans and rhesus macaque monkeys. Using high-throughput data collection systems for human and monkey psychophysics, we collected over one million behavioral trials for 2400 images of 24 broadly sampled basic-level objects, resulting in 276 binary object discrimination tasks. Consistent with previous work, we observed that state-of-the-art deep, feed-forward, convolutional ANNs trained for visual categorization (termed $\mathrm{DCNN}_{\mathrm{IC}}$ models) accurately predicted primate patterns of object-level confusion (e.g. how often a camel is confused with a dog, on average). However, when we examined behavioral performance for individual images within each object discrimination task, we found that all of the $\mathrm{DCNN}_{\mathrm{IC}}$ models were significantly non-predictive of primate performance. We found that this prediction failure was not accounted for by simple image attributes, nor was it rescued by simple model modifications. These results show that current $\mathrm{DCNN}_{\mathrm{IC}}$ models cannot account for the image-level behavioral patterns of primates, even when images are not optimized to be adversarial. This suggests that new ANN models are needed to more precisely capture the neural mechanisms underlying primate object vision, and that high-resolution, large-scale behavioral metrics could serve as a strong constraint for discovering such models.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SIGNIFICANCE STATEMENT

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, 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 those high-resolution metrics, we found that all tested ANN models significantly diverged from primate behavior. Going forward, these high-resolution, large-scale behavioral metrics could serve as a strong constraint for discovering better ANN models of the primate visual system.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, Zoccolan et al. 2012), a deep hierarchical neural network (NN) of visual cortical areas. Thus, a primary neuroscience goal is to construct computational models that quantitatively explain the mechanisms underlying this ability, i.e. to discover artificial neural networks (ANNs) that accurately predict neuronal firing rate responses at all levels of the ventral stream, as well as its behavioral output. With respect to this goal, specific models within a large family of deep, convolutional neural networks (DCNNs) have been put forth as the leading ANN models of the ventral stream (Yamins and DiCarlo 2016). Specifically, the best such models are DCNNs optimized by supervised training on large-scale category-labeled image-sets (ImageNet) to match human-level object categorization performance (Krizhevsky, Sutskever et al. 2012, LeCun, Bengio et al. 2015); we refer to this sub-family of DCNN models as $DCNN_{IC}$ models (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 that are highly similar to neuronal representations in mid (V4) and high level cortical (IT) areas of the primate ventral visual stream (Yamins, Hong et al. 2013, Cadieu, Hong et al. 2014, Khaligh-Razavi and Kriegeskorte 2014, Yamins, Hong et al. 2014), and they also exhibit output patterns that are remarkably similar to the behavioral patterns of pairwise object confusions of primates in the domain of basic-level core object recognition (Rajalingham, Schmidt et al. 2015). As such, $DCNN_{IC}$ models may provide a quantitative account of the neural mechanisms underlying primate core object recognition behavior.

However, several studies have shown that the $DCNN_{IC}$ models can diverge drastically from humans in object recognition behavior, especially with regards to particular images optimized to be adversarial to these networks (Goodfellow, Shlens et al. 2014, Nguyen, Yosinski et al. 2015). Recent work demonstrated that such adversarial images are likely not isolated instances, suggesting that $DCNN_{IC}$ models may not match humans across larger image domains

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Goodfellow, Shlens et al. 2014). Related work has shown that specific image distortions (e.g. adding noise, blurring, inverting) are disproportionately challenging to current DCNNs, as compared to humans (Dodge and Karam 2017, Geirhos, Janssen et al. 2017, Hosseini, Xiao et al. 2017). Such image-specific failures of the current ANN models would likely not be captured by object-level behavioral metrics, such as the pattern of pairwise object confusions mentioned above (Rajalingham, Schmidt et al. 2015), that are computed by pooling over hundreds of images and thus are not sensitive to the fact that some images of an object are more challenging than other images of the same object. That limitation of prior work is due largely to data scale: reliable behavioral performance estimation requires many (20+) repeated measurements to assess behavioral discriminability per experimental condition, and large-scale measurements at the image-level are comprised of many such conditions (e.g. 2400 images with 23 distractor choices per image results in 55200 conditions for measuring discrimination performance). To overcome this limitation of prior work, we expanded the scale of our data collection to approximately 1.8 million trials from humans and monkeys, and we developed new behavioral metrics to reliably measure and characterize behavior at the resolution of images. Here, we directly compared leading DCNN models to primates—human and rhesus macaque monkeys—over the domain of core object recognition behavior at the high resolution of individual images.

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 and Cox 2007, DiCarlo, Zoccolan et al. 2012), operationalized as images of high view uncertainty presented in the central 10° of the visual field for durations under 200ms. For this study, we further restricted our sampled object discrimination tests within that domain to “basic-level” object discriminations, as defined previously (Rosch, Mervis et al. 1976), and to rigid object transformations. Within this domain, we collected over a million behavioral trials to make large-scale, high-resolution measurements of human and monkey behavior 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, Schmidt et al. 2015). However, at the high resolution of individual images, we found that the behavior of each and

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

every one of the $\mathrm{DCNN}_{\mathrm{IC}}$ models was significantly different from human and monkey behavior. This model prediction failure could not be easily rescued by modifications, such as primate-like retinal input sampling or additional model training. Taken together, these results show that current $\mathrm{DCNN}_{\mathrm{IC}}$ models do not fully account for the image-level behavioral patterns of primates, even when images are not optimized to be adversarial, 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 metrics such as those produced here could serve as a strong top-down constraint for efficiently discovering such models.

MATERIALS & METHODS

Visual images

We examined basic-level, core object recognition behavior using a set of 24 broadly-sampled objects that we previously found to be highly reliably labeled by independent human subjects, based on the definition of basic-level proposed by (Rosch, Mervis et al. 1976). For each object, we generated 100 naturalistic synthetic images 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 background. To do this, each object was first assigned a canonical position (center of gaze), scale (~2 degrees) and pose, and then its viewing parameters were randomly sampled uniformly from the following ranges for object translation ([-3,3] degrees in both h and v), rotation ([-180,180] degrees in all three axes) and scale ([x0.7, x1.7]. Backgrounds 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 visual recognition system (human, animal or model) to tackle the “invariance problem,” the computational crux of object recognition (Ullman and Humphreys 1996, Pinto, Cox et al. 2008). Using this procedure, we previously generated 2400 images (100 images per object) rendered at 1024x1024 pixel resolution with 256-level gray scale and subsequently resized to 256x256 pixel resolution for human psychophysics, monkey psychophysics and model evaluation (Rajalingham, Schmidt et al. 2015). In the current work, we focused our analyses on a randomly subsampled, and then

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fixed, sub-set of 240 images (10 images per object; here referred to as the "primary test images"). Figure 1A 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 viewpoint attributes as well as its low-level image attributes (i.e. pixel-wise distributional statistics computed from the final rendered image). Viewpoint attributes consisted of size, eccentricity and relative pose of the object in the image. For each synthetic object, we first defined its "canonical" 3D pose vector, based on independent human judgments. To compute the relative pose (RP) 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 (rxy, rxz, ryz) 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). Low-level image attributes included mean luminance of the image, segmentation index of the object from the background in the image, and spatial frequency content of the image background. The mean luminance was computed as the mean of all pixel intensities for each image. To compute the segmentation index, we measured the absolute difference in intensity 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). To compute an attribute characterizing the background spatial frequency (BSF), we first converted each image's background (prior to placing the foreground object) into the frequency domain using a 2D FFT, which we summarized using the spectral centroid. Figure 5C shows example images with varying attribute values for the three viewpoint attributes and the three low-level attributes.

Core object recognition behavioral paradigm

As in our previous work (Rajalingham, Schmidt 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.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 this task paradigm. 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 1B 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° of visual angle in size, and the choice object images were presented at ±6–8° of eccentricity along the horizontal meridian.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We measured human behavior using the online Amazon MTurk platform (see Figure 1C), 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, Hong et al. 2015, Rajalingham, Schmidt 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 (~xx) 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±70 trials/image on this subsampled set, versus 271±93 trials/image on the remaining images, mean ± 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±2.0 trials/image per subject, mean ± SD), in order to mitigate potential alternative behavioral strategies (e.g. "memorization" of images) that could potentially arise from a finite image set. Behavioral metrics at the object-level (B.O1, B.O2, see Behavioral Metrics) were measured using all 2400 test images, while image-level behavioral metrics (B.I1n, B.I2n) were measured using the 240 primary test images. (We observed qualitatively similar results for those metrics using 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).

Four other human subjects were separately recruited on MTurk to each perform a large number of trials on the same images and tasks (53,097±15,278 trials/subject, mean ± SD). Behavioral data from these four 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. This held-out human pool serves to provide a "gold-standard" for benchmarking all other candidate models.

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

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 ~2,000 trials per day per monkey, we collected a total of 836,117 trials from the five monkey subjects over a period of ~3 months.

Monkey training is described in detail elsewhere (Rajalingham, Schmidt 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 pattern of the first and second half of trials were highly consistent to each other (B.I1 consistency of 0.85±0.06, mean ± SEM over five monkey subjects), 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 1B), with the exception that trials were initiated by touching a white "fixation" circle horizontally

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 ~6 degrees of visual angle, 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. And the relatively short (but natural and high performing (Cadieu, Hong et al. 2014)) test image duration (100 ms) insured that saccadic eye movements were unlike to influence test image performance (as they generally take ~200 ms to initiate in response to the test image, and thus well after the test image has been extinguished).

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 benchmark over the past five years: AlexNet (Krizhevsky, Sutskever et al. 2012), NYU (Zeiler and Fergus 2014), VGG (Simonyan and Zisserman 2014), GoogleNet (Szegedy, Zaremba et al. 2013), Resnet (He, Zhang et al. 2016), and Inception-v3 (Szegedy, Zaremba et al. 2013). As this is only a subset of possible DCNN models, we refer to these as the $\mathrm{DCNN}_{\mathrm{IC}}$ (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. The final feature layer of ImageNet trained $\mathrm{DCNN}_{\mathrm{IC}}$ models corresponds to the probability output of this 1000-way classification task. We adapted these ImageNet-trained models to our 24-way object recognition task by re-training the final class probability layer, while holding all other layers fixed. In practice, this was done by extracting features from the penultimate layer of each $\mathrm{DCNN}_{\mathrm{IC}}$ (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 estimate the output class probability for each image. This procedure is illustrated in Figure 1C. 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* below).

From these analyses, we selected the most consistent $\mathrm{DCNN}_{\mathrm{IC}}$ architecture (Inception-v3), 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

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. 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. 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).

Behavioral metrics

We measured the object recognition behavior of humans, macaques and DCNN$_{IC}$ models using many test images in 276 interleaved binary object discrimination tasks (see above) To analyze these behavioral data, we here introduce four behavioral (B) metrics of increasing richness, but requiring increasing amounts of data to measure reliably. Each behavioral metric computes a pattern of unbiased behavioral performance, using a sensitivity index: $d' = Z(HitRate) - Z(FalseAlarmRate)$, 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 four behavioral metrics, varying the hit-rate resolution (image-level or object-level) and the false-alarm resolution (one-versus-all or one-versus-other). Briefly, 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 B.O1 metric measured here has 24 independent values. The one-versus-other object-level performance metric (termed B.O2) estimates the discriminability of

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Naturally, object-level and image-level behavioral patterns 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 the B.O1 and B.O2 metrics), we estimated normalized image-level behavioral metrics by subtracting the mean performance values over all images of the same object and task. This process is schematically illustrated in Figure 3A. We focus on these normalized image-level behavioral metrics (termed B.I1n, B.I2n) for image-level comparisons between models and primates (see Results).

Behavioral Consistency

For each visual system, we randomly split all behavioral trials into two equal halves and computed each behavioral metric on each half. To estimate the reliability of each system's behavioral pattern given the amount of data collected, we computed the Pearson correlation between behavioral patterns estimated from separate halves of the data (random split-halves of trials). To quantify the similarity between a model visual system and the human visual system, we use a measure called the noise-adjusted human "consistency" (referred to in the text as "human consistency") as previously defined (Johnson, Hsiao et al. 2002). Consistency ($\tilde{\rho}$) is computed for each of the four behavioral metrics. Specifically, for each metric, we computed the

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pearson correlation over all the independent measurements in the metric from the model ($\mathbf{m}$) and the human ($\mathbf{h}$), and we then normalize that raw Pearson correlation by the geometric mean of the split-half internal reliability of the same behavioral metric measured for each system: $\tilde{\rho}(\boldsymbol{m},\boldsymbol{h}) = \frac{\rho(\boldsymbol{m},\boldsymbol{h})}{\sqrt{\rho(\boldsymbol{m},\boldsymbol{m})\rho(\boldsymbol{h},\boldsymbol{h})}}$.

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 noise-adjusted correlation measure for consistency was to account for variance in the behavioral patterns 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 (Johnson, Hsiao et al. 2002).

Characterization of Residuals

In addition to measuring the similarity between the behavioral patterns of primates and models (using consistency analyses, as described above), we examined the corresponding differences, or "residual behavioral patterns." Each candidate visual system model's residual behavioral pattern was estimated as the residual of a linear least squares regression on the human pool data (one behavioral performance value per test image, thus 240 values) and we included a free intercept parameter. This procedure effectively captures the differences between human and model behavior after accounting for overall performance differences. Residual patterns were estimated on disjoint split-halves of trials, repeating 10 times with random trial permutations. We focused on the normalized one-versus-all image-level performance pattern (B.I1n) to reliably measure image-level differences between primates and models as that 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 humans is reliably shared across different models, we measured the Pearson correlation between the residual patterns of pairs of models. Residual similarity was quantified as the proportion of

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

shared variance, defined as the square of the noise-adjusted correlation between residual patterns (the noise-adjustment was done as defined in equation above). Correlations of residual patterns 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 consistency 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 behavioral patterns 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 four subjects); difficult images were defined as ones with performance below the 50th and 25th percentiles 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 pattern from image attributes, including viewpoint attributes (e.g. object size, eccentricity, pose) and pixel attributes (e.g. mean luminance, background spatial frequency, segmentation-index). Briefly, we constructed a design matrix from the image attributes (using individual attributes, groups of attributes, or all attributes), and used multiple linear least squares regression to predict the image-level residual pattern. 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 pattern.

Primate zone

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (i.e. have consistency of 1.0) with this archetypal human. 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 consistency of the corresponding "archetypal monkey" to the human pool. Rather, we indirectly estimated this consistency by first measuring consistency as a function of number of individual subjects pooled together (n), and extrapolating the 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 $\tilde{\rho}(n) = a + b \cdot e^{-cn}$ (see Figure 4).

For each behavioral metric, we defined a "primate zone" as the range of consistency values delimited by consistency 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 consistency relative to the human pool of a large (i.e. infinitely many subjects) pool of rhesus macaque monkeys; $\tilde{\rho}_{H\infty}$ is by definition equal to 1.0. Thus, the primate zone defines a range of 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 behavioral 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$_{\mathrm{IC}}$ 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$_{\mathrm{IC}}$ models is within the primate zone. To test this hypothesis, we estimate 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

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 $\tilde{\rho}_{M\infty}$ 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.

RESULTS

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 (B) 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 computational, was tested on the same 2400 images (24 objects, 100 images/object) in the same 276 interleaved binary object recognition tasks. Each system's behavior was characterized at multiple resolutions (see *Behavioral metrics* in Methods) and directly compared to the corresponding behavioral metric of the archetypal human (defined as the average behavior over a large pool of human subjects tested; see Methods). The overarching logic of this study is that, if two visual systems are equivalent, they should produce statistically indistinguishable behavioral metrics (B).

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 metric vector is 24 dimensional. Figure 2A shows the B.O1 metric vector 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 bin

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 pattern'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 2B directly compares the B.O1 metric vector 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 BO1 metric vector. We observe a tighter correspondence to the human behavioral pattern for the DCNN_{IC} model visual system than for the baseline pixel model visual systems. We quantified that similarity using a noise-adjusted correlation between each pair of B.O1 vectors (termed *consistency,* following (Johnson, Hsiao 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 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 four human subjects (black dot) and a pool of five macaque monkey subjects (gray dot), and we observed that both yielded B.O1 vectors that were highly consistent to the human pool ($\tilde{\rho} = 0.90, 0.97$ for monkey pool and held-out human pool, respectively). We defined a range of consistency values, termed the "primate zone" (shaded gray area), delimited by extrapolated human consistency estimates of large pools of macaques and humans (see Methods, Figure 4). With respect to the B.O1 metric, all tested DCNN_{IC} visual system models were either within or very close to this zone, while the baseline pixel visual system model and the low-level V1 visual system model were not ($\tilde{\rho} = 0.40, 0.67$ for pixels and V1 models, respectively). Based on the B.O1 behavioral metric alone, the hypothesis that the human consistency of DCNN_{IC} models is within the primate zone could not be rejected (p = 0.54, exact test, see Methods).

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 yields a symmetric matrix that is referred to here as the B.O2 metric. Figure 2D shows the B.O2 metric for pooled human, pooled monkey,

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 pattern'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 2E, which shows the consistency relative to the human pool 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 consistent to the human pool 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 consistent with the human pool ($\tilde{\rho} > 0.8$) 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. Again, based on the B.O2 behavioral metric, the hypothesis that the human consistency of the DCNN_{IC} models is within the primate zone could not be rejected (p = 0.99, exact test).

Taken together, humans, monkeys, and current DCNN_{IC} models all share similar patterns of object-level behavioral performance patterns (B.O1 and B.O2 metrics) 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 pattern of performances 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 image-level performance. Note that, from the point of view of the subjects, the behavioral tasks are identical to those already described. We are simply aiming to measure and compare their patterns of performance at much higher resolution.

Image-level behavioral comparison

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 metric, we focused our analyses on normalized image-level performance patterns. This normalization procedure is schematically illustrated in Figure 3A for the one-versus-all image-level performance pattern (240-dimensional, 10 images/object) to obtain the normalized one-versus-all image-level behavioral metric (termed B.I1n metric, see Methods). Figure 3B shows the B.I1n metric for the pooled human, pooled monkey, and several $\mathrm{DCNN}_{\mathrm{IC}}$ 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 3A), where warmer colors indicate lower values; color scales are not shown but span each pattern's full range. Figure 3D shows the consistency to the human pool with respect to the B.I1n metric for all tested models. Unlike with object-level behavioral metrics, we now observe a divergence between $\mathrm{DCNN}_{\mathrm{IC}}$ models and primates. Both the monkey pool and the held-out human pool remain highly consistent with the pooled human with respect to this metric ($\tilde{\rho} = 0.77, 0.96$ for monkeys, humans respectively), but all $\mathrm{DCNN}_{\mathrm{IC}}$ models were significantly less consistent (Inception-v3: $\tilde{\rho} = 0.62$) and well outside of the defined "primate zone" of I1_c consistency to the human pool. Indeed, based on the B.I1n behavioral metric, the hypothesis that the human consistency of $\mathrm{DCNN}_{\mathrm{IC}}$ models is within the primate zone is strongly rejected ($p = 6.16e\text{-}8$, exact test, see Methods).

We can zoom in further on this metric 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 3D shows the B.I2n metric as 240x24 matrices for the pooled human, pooled monkey and top $\mathrm{DCNN}_{\mathrm{IC}}$ visual system models.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Color scales are not shown but span each pattern'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 3E shows the human consistency with respect to the B.I2n metric for all tested visual system models. Extending our observations using the vector of image difficulties (B.I1n), we observe a similar divergence between primates and $\mathrm{DCNN_{IC}}$ visual system models on the matrix pattern of image-by-distractor difficulties (I2n). Specifically, both the monkey pool and held-out human pool remain highly consistent with the pooled human ($\tilde{\rho}$ = 0.75, 0.77 for monkeys, humans respectively), while all tested $\mathrm{DCNN_{IC}}$ models are significantly less consistent (Inception-v3: $\tilde{\rho}$ = 0.53) falling well outside of the defined "primate zone" of I2n consistency to the human pool. Once again, based on the B.I2n behavioral metric, the hypothesis that the human consistency of $\mathrm{DCNN_{IC}}$ models is within the primate zone is strongly rejected (p = 3.17e-18, exact test, see Methods).

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 consistency 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 the human (pool) 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

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 4A shows the measured human consistency for each of the four behavioral metrics, for 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 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 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 4A, 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 4A, last two panels).

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

Next, we wondered if some relatively simple changes to the $\mathrm{DCNN_{IC}}$ 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 $\mathrm{DCNN_{IC}}$ model visual system that scored the highest in our benchmarks (Inception-v3): we (1) changed the input to the model to be more primate-like in its

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

retinal sampling (Inception-v3 + retina-like), (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), we found that none of these modifications led to a significant improvement in its human consistency on the behavioral metrics (Figure 4B). Thus, the failure of current $\mathrm{DCNN}_{\mathrm{IC}}$ models to accurately capture the image-level behavioral patterns of primates cannot be rescued by simple modifications on a fixed architecture.

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

Taken together, Figures 2, 3 and 4 suggest that current $\mathrm{DCNN}_{\mathrm{IC}}$ visual system models fail to accurately capture the image-level behavioral patterns of humans and monkeys. To further examine this failure in the hopes of providing clues for model improvement, we examined the residual image-level behavioral patterns of all the visual system models, relative to the pooled human. For each model, we computed its residual image-level behavioral pattern as the difference (positive or negative) of a linear least squares regression of the model predictions with the human pool observations. For this analysis, we focused on the B.I1n metric as it showed a clear divergence of $\mathrm{DCNN}_{\mathrm{IC}}$ 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 image-level behavioral patterns are shared between different visual system models.

Figure 5A shows the similarity between the residual image-level patterns of all pairs of models; the color of bin (i,j) indicates the proportion of explainable variance that is shared between the residual image-level patterns 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

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"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 5B shows the mean (±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 (blue and red dots, selected based on held-out human data, see Methods). First, consistent with the results shown in Figure 3, we note that the residual image-level patterns 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 image-level residual patterns were highly reliable between different instances of this fixed $DCNN_{IC}$ model, across random training initializations (region 2: r^2=0.79, 0.77 for all and most difficult images, respectively). Interestingly, residual patterns 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 patterns 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 image-level behavioral pattern of the Inception-v3 architecture on a range of image attributes. Specifically, we considered both object viewpoint attributes (the size, eccentricity, and pose of the object) and pixel attributes (mean luminance, background spatial frequency, segmentation index) of each image. We used multivariate regressions to predict the residual pattern from

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

groups of several image attributes (e.g. from all attributes), and also considered each attribute individually using univariate regressions. Figure 6A shows example images (sampled from the full set of 2400 images) with increasing attribute value for each of these six image attributes. While the $\mathrm{DCNN}_{\mathrm{IC}}$ 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 6B). 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. The similarity of the patterns in Figure 6B between primates and the $\mathrm{DCNN}_{\mathrm{IC}}$ 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 $\mathrm{DCNN}_{\mathrm{IC}}$ models failed to capture a large portion of the image-by-image variation (Figure 3). In particular, Figure 6C shows the proportion of variance explained by specific image attributes for the residual, patterns of monkeys (dark gray), Inception-v3 models (dark blue), and all $\mathrm{DCNN}_{\mathrm{IC}}$ models (light blue). We found that, taken together, all six of these image attributes explained only ~10% of the variance in the image-wise residual between humans and $\mathrm{DCNN}_{\mathrm{IC}}$. Furthermore, we found that pixel attributes, rather than viewpoint attributes, contributed the majority of this explanatory power. Each individual attribute could explain at most a small amount of residual behavioral variance (<5% of the explainable variance). In sum, these analyses show that some behavioral effects that might provide intuitive clues to modify the $\mathrm{DCNN}_{\mathrm{IC}}$ 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 6C) argue that the $\mathrm{DCNN}_{\mathrm{IC}}$ visual system models' failure to capture primate image-level performance patterns cannot be further accounted for by these simple image attributes and likely stem from other factors.

DISCUSSION

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Broadly, our scientific goal is to discover computational 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, drawn from a large family of deep convolutional neural networks and optimized to achieve high levels of object categorization performance on large-scale image-sets, accurately capture the coarse behavioral patterns of primates in core object recognition tasks while the internal hidden neurons of those same models also predict a large fraction of primate ventral stream neural response variance to images (Cadieu, Hong et al. 2014, Khaligh-Razavi and Kriegeskorte 2014, Yamins, Hong et al. 2014, Güçlü and van Gerven 2015, Rajalingham, Schmidt et al. 2015, Kheradpisheh, Ghodrati et al. 2016, Kubilius, Bracci et al. 2016). 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 patterns at both object and image level resolution in the domain of core object recognition. Our primary novel result is that leading DCNN_{IC} models fail to fully replicate the image-level behavioral patterns 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.

While it had previously been shown that DCNN_{IC} models can diverge from human behavior on specifically chosen adversarial images (Szegedy, Zaremba 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. Furthermore, we showed that the failure of current DCNN_{IC} models to accurately predict primate behavioral patterns cannot be explained by simple image attributes (e.g. object viewpoint meta-parameters and low-level image statistics) and cannot be rescued by simple model modifications (input image sampling, model training, and classifier variations). Taken together, these results expose a general failure of current DCNN_{IC} to fully replicate the image-level behavioral patterns of primates and suggest that new ANN models are needed to more precisely capture the neural mechanisms underlying primate object vision.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

With regards to new ANN models, we can attempt to make prospective inferences about new and untested models from the data presented here. Based on the observed distribution of image-level behavioral consistency values for the tested $DCNN_{IC}$ models, one could infer that yet untested model instances sampled identically (i.e. from the same model sub-family) are highly likely to have similarly inadequate image-level behavioral consistency with primates. While we cannot rule out the possibility that at least one model instance within the $DCNN_{IC}$ class fully matches image-level human patterns, the probability of sampling such a model is vanishingly small ($p<10^{-18}$ for B.I2n consistency, estimated using exact test using Gaussian kernel density estimation, see Methods, Results). An important caveat of this inference is that we may have poorly estimated the consistency distribution, as we did not exhaustively sample this model 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 consistency of new (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). Thus, following this trend, it is possible that the evolution of "next-generation" models within the $DCNN_{IC}$ sub-family could meet the criterion for success of primate-like behavior.

Alternatively, it is possible that new $DCNN_{IC}$ models would 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 individual changes to the $DCNN_{IC}$ models—each creating a new ANN model sub-family—may be the best way forward, as $DCNN_{IC}$ models currently best explain both the behavioral and neural phenomena of core object recognition.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To reach that goal of finding a new ANN model sub-family that is an even better mechanistic model of the primate ventral visual stream, we propose that even larger-scale, high-resolution behavioral measurements than previously used, such as expanded versions of the patterns of image-level performance presented here, could serve as a useful top-down optimization constraint. Not only do these high-resolution behavioral metrics 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 constraint 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 ~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, Hong 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 refrain from making strong inferences from small behavioral differences, such as the 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 the large-scale, high-resolution behavioral datasets that could be efficiently collected in both humans and monkeys, specifically toward the goal of constraining future models of the primate ventral visual stream.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Cadieu, C. F., et al. (2014). "Deep neural networks rival the representation of primate IT cortex for core visual object recognition." PLoS computational biology **10**(12): e1003963.

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

DiCarlo, J. J., et al. (2012). "How does the brain solve visual object recognition?" Neuron **73**(3): 415-434.

Dodge, S. and L. Karam (2017). "A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions." arXiv preprint arXiv:1705.02498.

Geirhos, R., et al. (2017). "Comparing deep neural networks against humans: object recognition when the signal gets weaker." arXiv preprint arXiv:1706.06969.

Goodfellow, I. J., et al. (2014). "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572.

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

He, K., et al. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.

Hosseini, H., et al. (2017). "On the Limitation of Convolutional Neural Networks in Recognizing Negative Images." human performance **4**(5): 6.

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

Johnson, K. O., et al. (2002). "Neural coding and the basic law of psychophysics." The Neuroscientist **8**(2): 111-121.

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

Kheradpisheh, S. R., et al. (2016). "Deep networks can resemble human feed-forward vision in invariant object recognition." Scientific reports **6**: 32672.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Krizhevsky, A., et al. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems.

Kubilius, J., et al. (2016). "Deep neural networks as a computational model for human shape sensitivity." PLoS computational biology **12**(4): e1004896.

LeCun, Y., et al. (2015). "Deep learning." Nature **521**(7553): 436-444.

Majaj, N. J., et al. (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.

Nguyen, A., et al. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Pinto, N., et al. (2008). "Why is real-world visual object recognition hard?" PLoS computational biology **4**(1): e27.

Rajalingham, R., et al. (2015). "Comparison of Object Recognition Behavior in Human and Monkey." The Journal of Neuroscience **35**(35): 12127-12136.

Rosch, E., et al. (1976). "Basic objects in natural categories." Cognitive psychology **8**(3): 382-439.

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

Szegedy, C., et al. (2013). "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199.

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

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

Yamins, D. L., et al. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. Advances in neural information processing systems.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Yamins, D. L., et al. (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences: 201403112.

Zeiler, M. D. and R. Fergus (2014). Visualizing and understanding convolutional networks. Computer Vision–ECCV 2014, Springer: 818-833.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLES

Table 1

Behavioral Metric	Hit Rate	False Alarm Rate
One-versus all object-level performance (B.O1) (24 x 1) $O_1(i) = Z\big(HR(i)\big) - Z\big(FAR(i)\big),$ $i = 1,2,\dots,24$	Proportion of trials when images of object *i* were correctly labeled as object i.	Proportion of trials when any image was incorrectly labeled as object i.
One-versus-other object-level performance B.O2 (24 x 24) $O_2(i,j) = Z\big(HR(i,j)\big) - Z\big(FAR(i,j)\big),$ $i = 1,2,\dots,24$ $j = 1,2,\dots,24$	Proportion of trials when images of object *i* were correctly labeled as i, when presented against distractor object *j*.	Proportion of trials when images of object *j* were incorrectly labeled as object *i*
One-versus-all image-level performance B.I1 (240 x 1) $I_1(ii) = Z\big(HR(ii)\big) - Z\big(FAR(ii)\big),$ $ii = 1,2,\dots,240$ $j = 1,2,\dots,24$	Proportion of trials when image ii was correctly classified as object i.	Proportion of trials when any image was incorrectly labeled as object i.
One-versus-other image-level performance B.I2 (240 x 24) $I_2(ii,j) = Z\big(HR(ii,j)\big) - Z\big(FAR(ii,j)\big),$ $ii = 1,2,\dots,240$ $j = 1,2,\dots,24$	Proportion of trials when image ii was correctly classified as object i, when presented against distractor object *j*.	Proportion of trials when images of object *j* were incorrectly labeled as object *i*

Table 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.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. Images and behavioral task. (A) Two (out of 100) example images for each of the 24 basic-level objects. To enforce true invariant object recognition behavior, we generated naturalistic synthetic images, each with one foreground object, by rendering a 3D model of each object with randomly chosen viewing parameters and placing that foreground object view onto a randomly chosen, natural image background. **(B)** Time course of example behavioral trial (zebra versus dog) for human psychophysics. Each trial initiated with a central fixation point for 500 ms, followed by 100 ms presentation of a square test image (spanning 6-8° of visual angle). After extinction of the test image, two choice images were shown to the left and right. Human participants were allowed to freely view the response images for up to 1000 ms and responded by clicking on one of the choice images; no feedback was given. To neutralize top-down feature attention, all 276 binary object discrimination tasks were randomly interleaved on a trial-by-trial basis. The monkey task paradigm was nearly identical to the human paradigm, with the exception that trials were initiated by touching a fixation circle horizontally centered on the bottom third of the screen, and successful trials were rewarded with juice while incorrect choices resulted in timeouts of 1–2.5s. **(C)** Large-scale and high-throughput psychophysics in humans (top left), monkeys (top right), and models (bottom). Human behavior was measured using the online Amazon MTurk platform, which enabled the rapid collection ~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. Deep convolutional neural network 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. All behavioral predictions of each visual system model were for images that were not seen in any phase of model training.

Figure 2. Object-level comparison to human behavior. (A) One-versus-all object-level performance (B.O1) metric for the pooled human (n=1472 human subjects), pooled monkey (n=5 monkey subjects), and several DCNN$_{IC}$ models. Each B.O1 pattern is shown as a 24-dimensional vector using a color scale; each colored bin corresponds to the system's

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

discriminability of one object against all others that were tested. The color scales span each pattern's full performance range, and warm colors indicate lower discriminability. **(B)** Direct comparison of the B.O1 metric vector computed from the behavioral output of a pixel visual system model (top panel) and a $\mathrm{DCNN}_{\mathrm{IC}}$ visual system model (Inception-v3, bottom panel) against that of the human B.O1 metric vector. **(C)** Consistency to the human pool, with respect to the B.O1 metric, for each of the tested model visual systems. The black and gray dots correspond to a held-out pool of four 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 consistency of a pool of infinitely many monkeys (see Figure 4A). **(D)** One-versus-other object-level performance (B.O2) metric for pooled human, pooled monkey, and several $\mathrm{DCNN}_{\mathrm{IC}}$ models. Each B.O2 pattern 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)** Consistency to the human pool, with respect to the B.O2 metric, for each of the tested model visual systems. Format is identical to (C).

Figure 3. Image-level comparison to human behavior. (A) Schematic for computing B.I1n metric. First, the one-versus-all image-level metric (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 metric (B.I1n) is estimated by subtracting the mean performance 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 performance (B.I1n) metric for the pooled human, pooled monkey, and several $\mathrm{DCNN}_{\mathrm{IC}}$ models. Each B.I1n pattern is shown as a 240-dimensional vector using a color scale, formatted as in (A). Color scales similar to Figure 2A. **(C)** Consistency to the human pool, with respect to the B.I1n metric, for each of the tested model visual systems. Format is identical to Figure 2C. **(D)** Normalized one-versus-other image-level performance (B.I2n) metric for pooled human, pooled monkey, and several $\mathrm{DCNN}_{\mathrm{IC}}$ models. Each B.I2n pattern 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*. Color scales similar to Figure 2A. **(E)** Consistency to the human pool, with respect to the B.I2n metric, for each of the tested model visual systems. Format is identical to Figure 2C.

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. 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. 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 DCNN_{IC} visual system model that scored the highest in our benchmarks (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.

Figure 5. 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 image-level behavioral patterns 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 Figure 5A, the similarity to the residuals of Inception-v3 (region 2 in (A)) is shown (mean ± SD, within each region) for all images (black dots), and for images that humans found to be particularly difficult (blue and red dots, selected based on held-out human data).

Figure 6. Dependence of primate and DCNN model behavior on object viewpoint and pixel attributes. (A) Example images with increasing attribute value, for each of the six pre-defined image attributes. **(B)** Dependence of performance as a function of six image attributes, for humans, monkeys and a DCNN_{IC} model (Inception-v3). **(C)** Proportion of explainable variance of the residual image-level behavioral pattern of monkeys (black), an Inception-v3 model (dark blue), and all DCNN_{IC} models (light blue) that is accounted for by each of the pre-defined image

bioRxiv preprint doi: https://doi.org/10.1101/240614; this version posted January 1, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

attributes. Error-bars correspond to SD over trial re-sampling for monkeys, over different model “subjects” for Inception-v3, and over different $\mathrm{DCNN}_{\mathrm{IC}}$ models for “Models (all)”.

Figure 1

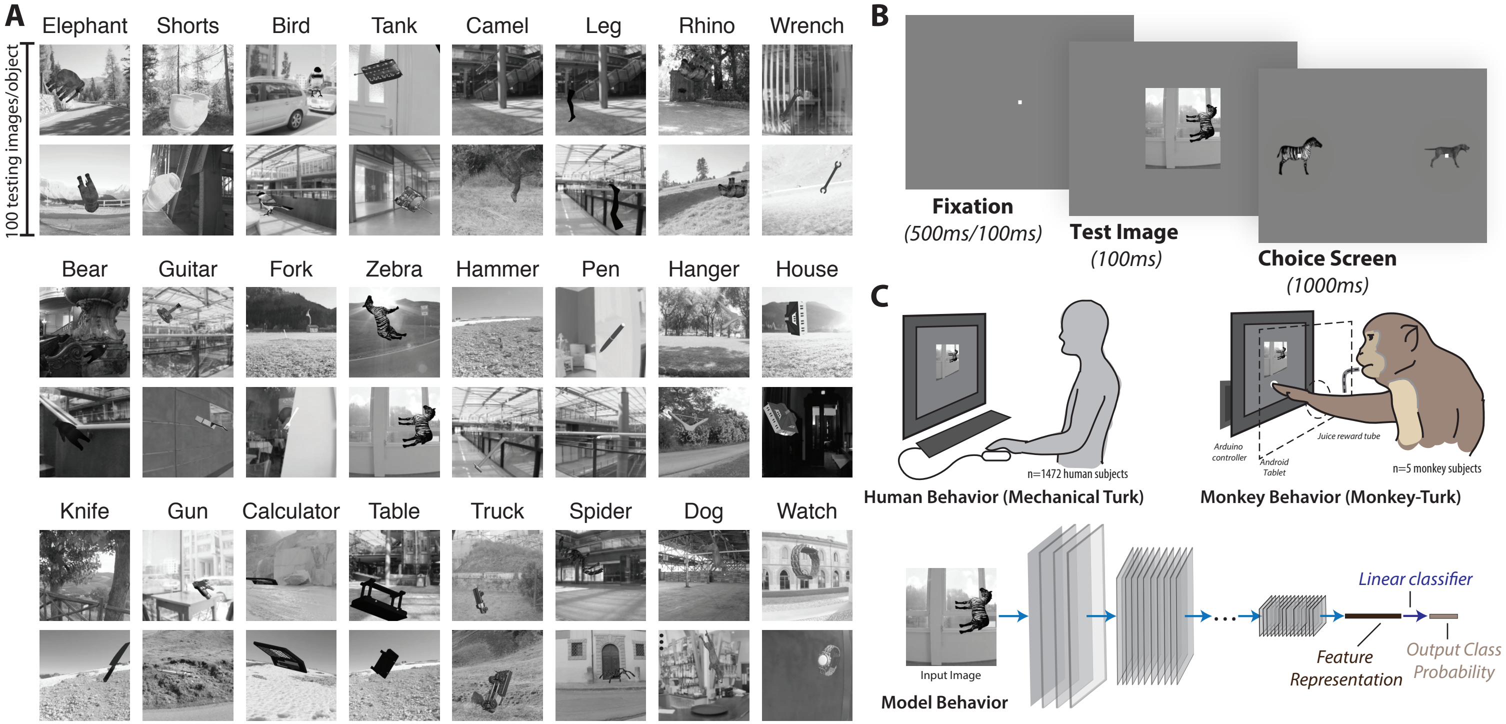

A
100 testing images/object
Elephant
Shorts
Bird
Tank
Camel
Leg
Rhino
Wrench
Bear
Guitar
Fork
Zebra
Hammer
Pen
Hanger
House
Knife
Gun
Calculator
Table
Truck
Spider
Dog
Watch
B
Fixation
(500ms/100ms)
Test Image
(100ms)
Choice Screen
(1000ms)
C
n=1472 human subjects
Human Behavior (Mechanical Turk)
Juice reward tube
Arduino controller
Android Tablet
n=5 monkey subjects
Monkey Behavior (Monkey-Turk)
Input Image
Model Behavior
Linear classifier
Feature Representation
Output Class Probability

Figure 2

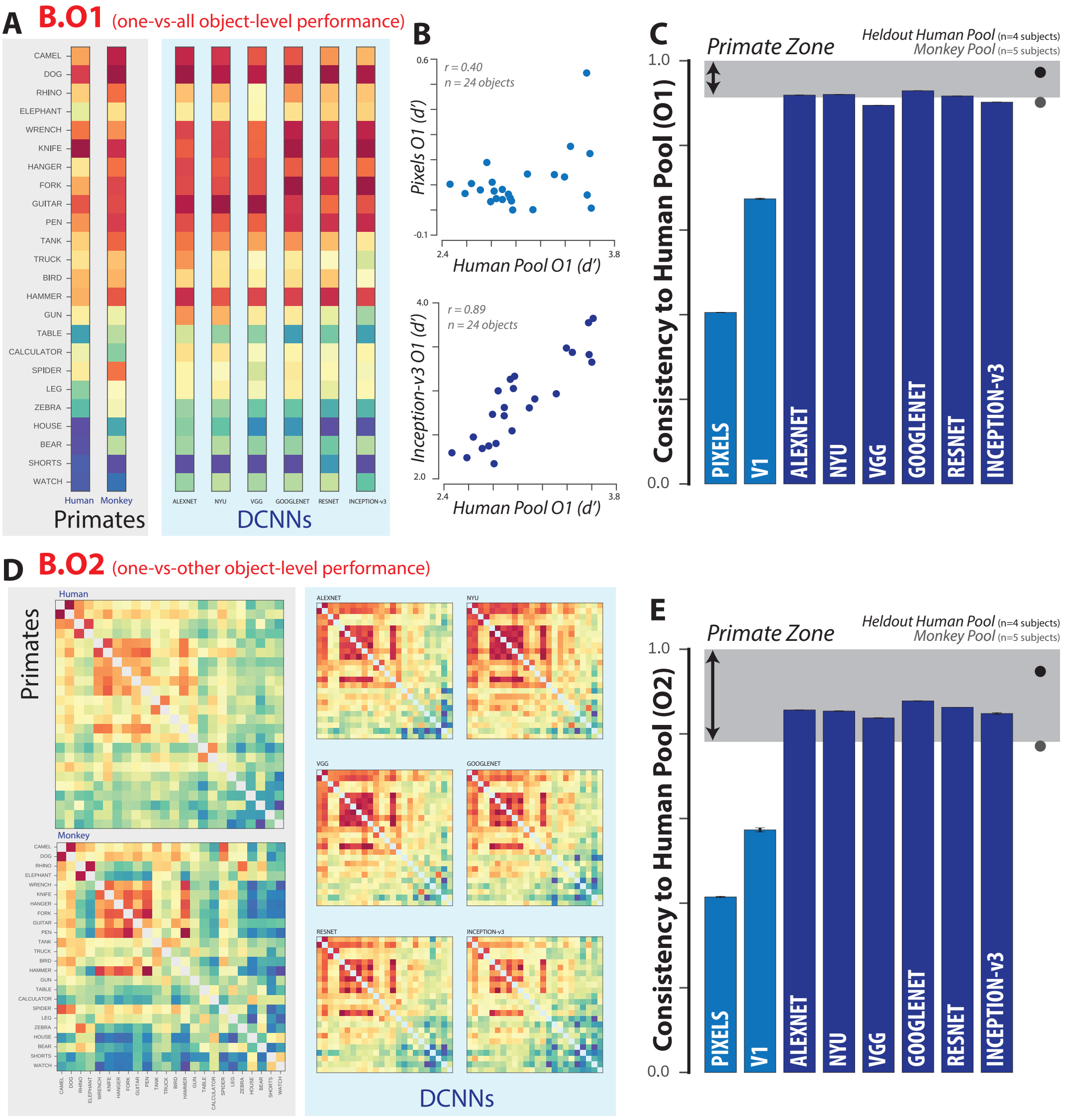

A
B.O1 (one-vs-all object-level performance)
CAMEL
DOG
RHINO
ELEPHANT
WRENCH
KNIFE
HANGER
FORK
GUITAR
PEN
TANK
TRUCK
BIRD
HAMMER
GUN
TABLE
CALCULATOR
SPIDER
LEG
ZEBRA
HOUSE
BEAR
SHORTS
WATCH
Human
Monkey
Primates
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3
DCNNs
B
0.6
-0.1
r = 0.40
n = 24 objects
Pixels O1 (d')
2.4
3.8
Human Pool O1 (d')
4.0
2.0
r = 0.89
n = 24 objects
Inception-v3 O1 (d')
2.4
3.8
Human Pool O1 (d')
C
Primate Zone
Heldout Human Pool (n=4 subjects)
Monkey Pool (n=5 subjects)
1.0
0.0
Consistency to Human Pool (O1)
PIXELS
V1
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3
D
B.O2 (one-vs-other object-level performance)
Primates
Human
Monkey
CAMEL
DOG
RHINO
ELEPHANT
WRENCH
KNIFE
HANGER
FORK
GUITAR
PEN
TANK
TRUCK
BIRD
HAMMER
GUN
TABLE
CALCULATOR
SPIDER
LEG
ZEBRA
HOUSE
BEAR
SHORTS
WATCH
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3
DCNNs
E
Primate Zone
Heldout Human Pool (n=4 subjects)
Monkey Pool (n=5 subjects)
1.0
0.0
Consistency to Human Pool (O2)
PIXELS
V1
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3

Figure 3

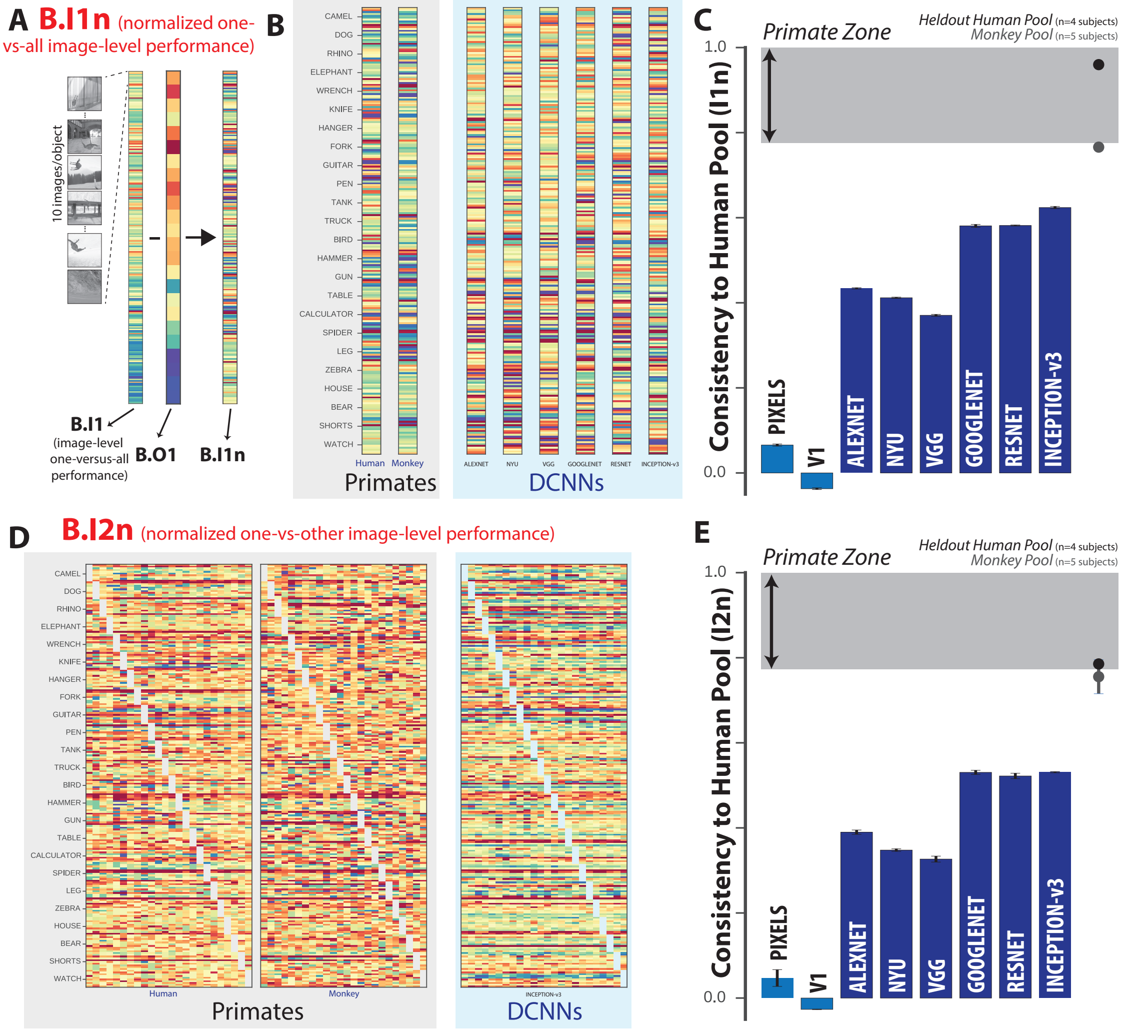

A
B.I1n (normalized one-vs-all image-level performance)
10 images/object
B.I1
(image-level one-versus-all performance)
B.O1
B.I1n
B
CAMEL
DOG
RHINO
ELEPHANT
WRENCH
KNIFE
HANGER
FORK
GUITAR
PEN
TANK
TRUCK
BIRD
HAMMER
GUN
TABLE
CALCULATOR
SPIDER
LEG
ZEBRA
HOUSE
BEAR
SHORTS
WATCH
Human
Monkey
Primates
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3
DCNNs
C
Primate Zone
Heldout Human Pool (n=4 subjects)
Monkey Pool (n=5 subjects)
1.0
0.0
Consistency to Human Pool (I1n)
PIXELS
V1
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3
D
B.I2n (normalized one-vs-other image-level performance)
CAMEL
DOG
RHINO
ELEPHANT
WRENCH
KNIFE
HANGER
FORK
GUITAR
PEN
TANK
TRUCK
BIRD
HAMMER
GUN
TABLE
CALCULATOR
SPIDER
LEG
ZEBRA
HOUSE
BEAR
SHORTS
WATCH
Human
Monkey
Primates
INCEPTION-v3
DCNNs
E
Primate Zone
Heldout Human Pool (n=4 subjects)
Monkey Pool (n=5 subjects)
1.0
0.0
Consistency to Human Pool (I2n)
PIXELS
V1
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3

Figure 4

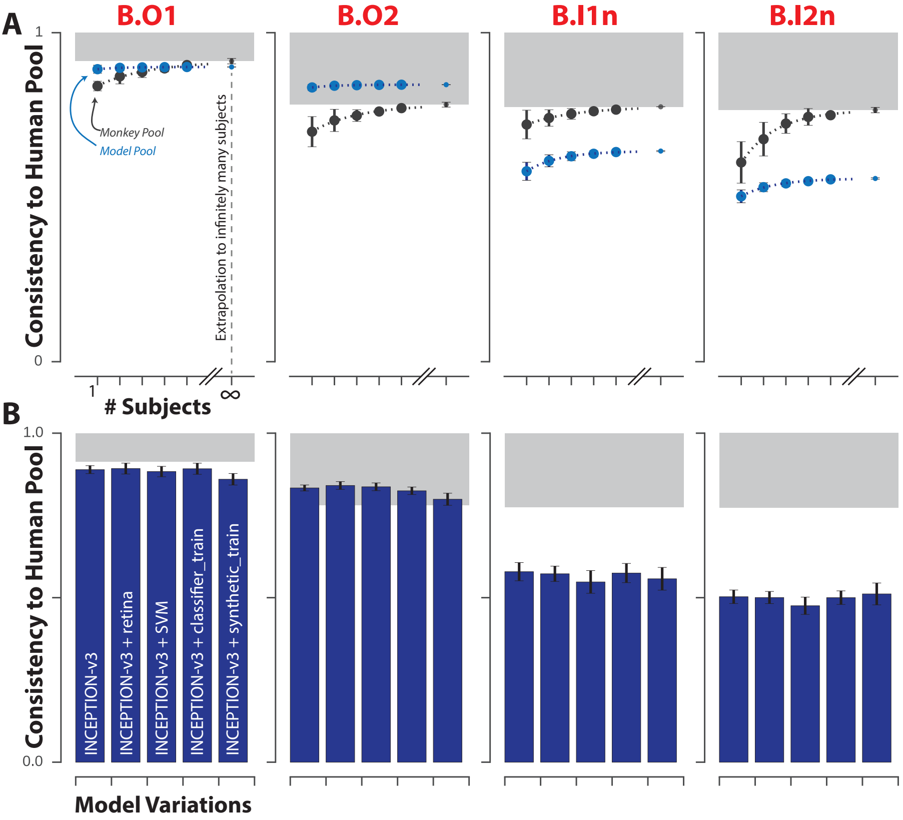

A
B.O1
B.O2
B.I1n
B.I2n
1
0
Consistency to Human Pool
Monkey Pool
Model Pool
Extrapolation to infinitely many subjects
1
∞
# Subjects
B
1.0
0.0
Consistency to Human Pool
INCEPTION-v3
INCEPTION-v3 + retina
INCEPTION-v3 + SVM
INCEPTION-v3 + classifier_train
INCEPTION-v3 + synthetic_train
Model Variations

Figure 5

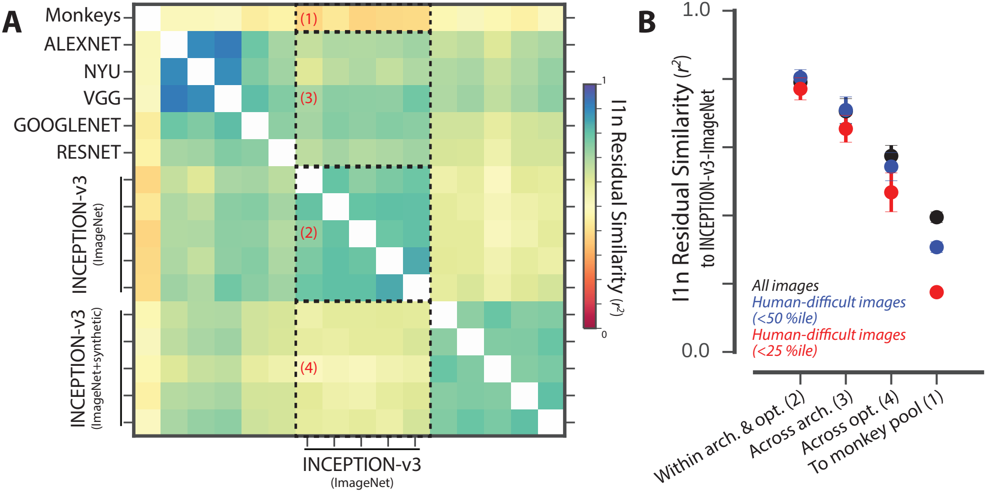
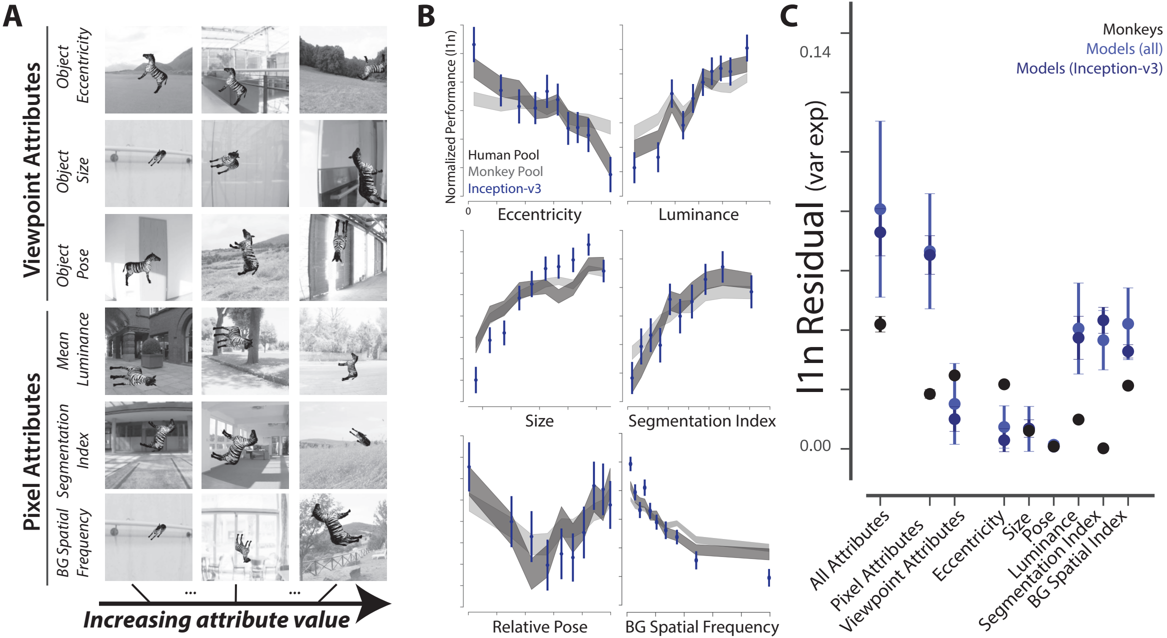

A
Monkeys
ALEXNET
NYU
VGG
GOOGLENET
RESNET
INCEPTION-v3
(ImageNet)
INCEPTION-v3
(ImageNet+synthetic)
(1)
(2)
(3)
(4)
INCEPTION-v3
(ImageNet)
1
0
I1n Residual Similarity (r^2)
B
1.0
0.0
I1n Residual Similarity (r^2)
to INCEPTION-v3-ImageNet
All images
Human-difficult images
(<50 %ile)
Human-difficult images
(<25 %ile)
Within arch. & opt. (2)
Across arch. (3)
Across opt. (4)
To monkey pool (1)

Figure 6



A
Viewpoint Attributes
Object Eccentricity
Object Size
Object Pose
Pixel Attributes
Mean Luminance
Segmentation Index
BG Spatial Frequency
...
...
Increasing attribute value
B
Normalized Performance (I1n)
0
Human Pool
Monkey Pool
Inception-v3
Eccentricity
Luminance
Size
Segmentation Index
Relative Pose
BG Spatial Frequency
C
Monkeys
Models (all)
Models (Inception-v3)
0.14
0.00
I1n Residual (var exp)
All Attributes
Pixel Attributes
Viewpoint Attributes
Eccentricity
Size
Pose
Luminance
Segmentation Index
BG Spatial Index