

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

4
5 Abbreviated title: Comparing object recognition behavior in humans, monkeys, and machines

6
7 Rishi Rajalingham*, Elias B. Issa*, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J.
8 DiCarlo

9
10 McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences
11 Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

12
13 *R.R. and E.B.I. contributed equally to this work.

14
15 Correspondence should be addressed to James J. DiCarlo, McGovern Institute for Brain
16 Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
17 77 Massachusetts Institute of Technology, 46-6161, Cambridge, MA 02139. E-mail:
18 dicarlo@mit.edu

19
20 E. Issa's present address: Department of Neuroscience, Zuckerman Mind Brain Behavior
21 Institute, Columbia University, New York, NY 10027

22
23

Targeting *Journal of Neuroscience*

24 Title 19 words

25 Abbreviated Title 57 characters

26 Abstract 291 words

27 Significance Statement 97 words

28 Introduction 895 words

29 Discussion 1051 words

30 Figures 6

31 *All word limits include citations

32
33

34 **AUTHOR CONTRIBUTIONS**

35 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
36 experiments. R.R., E.B.I., and P.B. performed the data analysis and modeling. R.R., E.B.I., and
37 J.J.D. wrote the manuscript.

38
39 **ACKNOWLEDGEMENTS**

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

43
44 **COMPETING FINANCIAL INTERESTS**

45 The authors declare no competing financial interests.

46

47 **ABSTRACT**

48

49 Primates—including humans—can typically recognize objects in visual images at a
50 glance, even in the face of naturally occurring identity preserving image transformations such as
51 changes in viewpoint. A primary neuroscience goal is to uncover neuron-level mechanistic
52 models that quantitatively explain this behavior, not only predicting average primate
53 performance, but also predicting primate performance for each and every image. Here, we
54 applied this stringent behavioral prediction test to the leading mechanistic models of primate
55 vision (specifically, deep, convolutional, artificial neural networks; ANNs) by directly
56 comparing their behavioral patterns, at high resolution over a large number of object
57 discrimination tasks, against those of humans and rhesus macaque monkeys. Using high-
58 throughput data collection systems for human and monkey psychophysics, we collected over one
59 million behavioral trials for 2400 images of 24 broadly sampled basic-level objects, resulting in
60 276 binary object discrimination tasks. Consistent with previous work, we observed that state-of-
61 the-art deep, feed-forward, convolutional ANNs trained for visual categorization (termed
62 $DCNN_{IC}$ models) accurately predicted primate patterns of object-level confusion (e.g. how often
63 a camel is confused with a dog, on average). However, when we examined behavioral
64 performance for individual images within each object discrimination task, we found that all of
65 the $DCNN_{IC}$ models were significantly non-predictive of primate performance. We found that
66 this prediction failure was not accounted for by simple image attributes, nor was it rescued by
67 simple model modifications. These results show that current $DCNN_{IC}$ models cannot account for
68 the image-level behavioral patterns of primates, even when images are not optimized to be
69 adversarial. This suggests that new ANN models are needed to more precisely capture the neural
70 mechanisms underlying primate object vision, and that high-resolution, large-scale behavioral
71 metrics could serve as a strong constraint for discovering such models.

72

73

74 **SIGNIFICANCE STATEMENT**

75

76 Recently, specific feed-forward deep convolutional artificial neural networks (ANNs)
77 models have dramatically advanced our quantitative understanding of the neural mechanisms
78 underlying primate core object recognition. In this work, we tested the limits of those ANNs by
79 systematically comparing the behavioral responses of these models with the behavioral responses
80 of humans and monkeys, at the resolution of individual images. Using those high-resolution
81 metrics, we found that all tested ANN models significantly diverged from primate behavior.
82 Going forward, these high-resolution, large-scale behavioral metrics could serve as a strong
83 constraint for discovering better ANN models of the primate visual system.

84

85 INTRODUCTION

86

87 Primates—both human and non-human—can typically recognize objects in visual images
88 at a glance, even in the face of naturally occurring identity-preserving transformations such as
89 changes in viewpoint. This view-invariant visual object recognition ability is thought to be
90 supported primarily by the primate ventral visual stream (DiCarlo, Zoccolan et al. 2012), a deep
91 hierarchical neural network (NN) of visual cortical areas. Thus, a primary neuroscience goal is to
92 construct computational models that quantitatively explain the mechanisms underlying this
93 ability, i.e. to discover artificial neural networks (ANNs) that accurately predict neuronal firing
94 rate responses at all levels of the ventral stream, as well as its behavioral output. With respect to
95 this goal, specific models within a large family of deep, convolutional neural networks (DCNNs)
96 have been put forth as the leading ANN models of the ventral stream (Yamins and DiCarlo
97 2016). Specifically, the best such models are DCNNs optimized by supervised training on large-
98 scale category-labeled image-sets (ImageNet) to match human-level object categorization
99 performance (Krizhevsky, Sutskever et al. 2012, LeCun, Bengio et al. 2015); we refer to this
100 sub-family of DCNN models as DCNN_{IC} models (to denote ImageNet-Categorization pre-
101 training), so as to distinguish them from all possible models in the DCNN family, and more
102 broadly, from the super-family of all ANNs. To date, it has been shown that DCNN_{IC} models
103 display internal feature representations that are highly similar to neuronal representations in mid
104 (V4) and high level cortical (IT) areas of the primate ventral visual stream (Yamins, Hong et al.
105 2013, Cadieu, Hong et al. 2014, Khaligh-Razavi and Kriegeskorte 2014, Yamins, Hong et al.
106 2014), and they also exhibit output patterns that are remarkably similar to the behavioral patterns
107 of pairwise object confusions of primates in the domain of basic-level core object recognition
108 (Rajalingham, Schmidt et al. 2015). As such, DCNN_{IC} models may provide a quantitative
109 account of the neural mechanisms underlying primate core object recognition behavior.

110

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

116 (Goodfellow, Shlens et al. 2014). Related work has shown that specific image distortions (e.g.
117 adding noise, blurring, inverting) are disproportionately challenging to current DCNNs, as
118 compared to humans (Dodge and Karam 2017, Geirhos, Janssen et al. 2017, Hosseini, Xiao et al.
119 2017). Such image-specific failures of the current ANN models would likely not be captured by
120 object-level behavioral metrics, such as the pattern of pairwise object confusions mentioned
121 above (Rajalingham, Schmidt et al. 2015), that are computed by pooling over hundreds of
122 images and thus are not sensitive to the fact that some images of an object are more challenging
123 than other images of the same object. That limitation of prior work is due largely to data scale:
124 reliable behavioral performance estimation requires many (20+) repeated measurements to assess
125 behavioral discriminability per experimental condition, and large-scale measurements at the
126 image-level are comprised of many such conditions (e.g. 2400 images with 23 distractor choices
127 per image results in 55200 conditions for measuring discrimination performance). To overcome
128 this limitation of prior work, we expanded the scale of our data collection to approximately 1.8
129 million trials from humans and monkeys, and we developed new behavioral metrics to reliably
130 measure and characterize behavior at the resolution of images. Here, we directly compared
131 leading DCNN models to primates—human and rhesus macaque monkeys—over the domain of
132 core object recognition behavior at the high resolution of individual images.

133
134 We focused on “core invariant object recognition”—the ability to identify objects in
135 visual images in the central visual field during a single, natural viewing fixation (DiCarlo and
136 Cox 2007, DiCarlo, Zoccolan et al. 2012), operationalized as images of high view uncertainty
137 presented in the central 10° of the visual field for durations under 200ms. For this study, we
138 further restricted our sampled object discrimination tests within that domain to “basic-level”
139 object discriminations, as defined previously (Rosch, Mervis et al. 1976), and to rigid object
140 transformations. Within this domain, we collected over a million behavioral trials to make large-
141 scale, high-resolution measurements of human and monkey behavior using high-throughput
142 psychophysical techniques—including a novel home-cage behavioral system for monkeys. These
143 data enabled us to systematically compare all systems at progressively higher resolution. At
144 lower resolutions, we replicated previous findings that humans, monkeys, and DCNN_{IC} models
145 all share a common pattern of object level confusion (Rajalingham, Schmidt et al. 2015).
146 However, at the high resolution of individual images, we found that the behavior of each and

147 every one of the DCNN_{IC} models was significantly different from human and monkey behavior.
148 This model prediction failure could not be easily rescued by modifications, such as primate-like
149 retinal input sampling or additional model training. Taken together, these results show that
150 current DCNN_{IC} models do not fully account for the image-level behavioral patterns of primates,
151 even when images are not optimized to be adversarial, suggesting that new ANN models are
152 needed to more precisely capture the neural mechanisms underlying primate object vision. To
153 this end, large-scale, high-resolution behavioral metrics such as those produced here could serve
154 as a strong top-down constraint for efficiently discovering such models.

155

156 **MATERIALS & METHODS**

157

158 *Visual images*

159 We examined basic-level, core object recognition behavior using a set of 24 broadly-
160 sampled objects that we previously found to be highly reliably labeled by independent human
161 subjects, based on the definition of basic-level proposed by (Rosch, Mervis et al. 1976). For each
162 object, we generated 100 naturalistic synthetic images by first rendering a 3D model of the object
163 with randomly chosen viewing parameters (2D position, 3D rotation and viewing distance), and
164 then placing that foreground object view onto a randomly chosen, natural image background. To
165 do this, each object was first assigned a canonical position (center of gaze), scale (~2 degrees)
166 and pose, and then its viewing parameters were randomly sampled uniformly from the following
167 ranges for object translation ([-3,3] degrees in both h and v), rotation ([-180,180] degrees in all
168 three axes) and scale ([x0.7, x1.7]). Backgrounds images were sampled randomly from a large
169 database of high-dynamic range images of indoor and outdoor scenes obtained from Dosch
170 Design (www.doschdesign.com). This image generation procedure enforces invariant object
171 recognition, rather than image matching, as it requires the visual recognition system (human,
172 animal or model) to tackle the “invariance problem,” the computational crux of object
173 recognition (Ullman and Humphreys 1996, Pinto, Cox et al. 2008). Using this procedure, we
174 previously generated 2400 images (100 images per object) rendered at 1024x1024 pixel
175 resolution with 256-level gray scale and subsequently resized to 256x256 pixel resolution for
176 human psychophysics, monkey psychophysics and model evaluation (Rajalingham, Schmidt et
177 al. 2015). In the current work, we focused our analyses on a randomly subsampled, and then

178 fixed, sub-set of 240 images (10 images per object; here referred to as the “primary test
179 images”). Figure 1A shows the full list of 24 objects, with two example images of each object.

180

181 Because all of the images were generated from synthetic 3D object models, we had
182 explicit knowledge of the viewpoint parameters (position, size, and pose) for each object in each
183 image, as well as perfect segmentation masks. Taking advantage of this feature, we characterized
184 each image based on these high-level viewpoint attributes as well as its low-level image
185 attributes (i.e. pixel-wise distributional statistics computed from the final rendered image).
186 Viewpoint attributes consisted of size, eccentricity and relative pose of the object in the image.
187 For each synthetic object, we first defined its “canonical” 3D pose vector, based on independent
188 human judgments. To compute the relative pose (RP) attribute of each image, we estimated the
189 difference between the object’s 3D pose and its canonical 3D pose. Pose differences were
190 computed as distances in unit quaternion representations: the 3D pose (rxy, rxz, ryz) was first
191 converted into unit quaternions, and distances between quaternions q_1, q_2 were estimated as
192 $\cos^{-1}|q_1 \cdot q_2|$ (Huynh 2009). Low-level image attributes included mean luminance of the
193 image, segmentation index of the object from the background in the image, and spatial frequency
194 content of the image background. The mean luminance was computed as the mean of all pixel
195 intensities for each image. To compute the segmentation index, we measured the absolute
196 difference in intensity between the mean of the pixel intensities corresponding to the object and
197 the mean of the background pixel intensities in the vicinity of the object (specifically, within 25
198 pixels of any object pixel, analogous to computing the local foreground-background luminance
199 difference of a foreground object in an image). To compute an attribute characterizing the
200 background spatial frequency (BSF), we first converted each image’s background (prior to
201 placing the foreground object) into the frequency domain using a 2D FFT, which we summarized
202 using the spectral centroid. Figure 5C shows example images with varying attribute values for
203 the three viewpoint attributes and the three low-level attributes.

204

205 *Core object recognition behavioral paradigm*

206 As in our previous work (Rajalingham, Schmidt et al. 2015), the behavioral task
207 paradigm consisted of a interleaved set of binary discrimination tasks. Each binary
208 discrimination task is an object discrimination task between a pair of objects (e.g. elephant vs.

209 bear). Each such binary task is balanced in that the test image is equally likely (50%) to be of
210 either of the two objects. On each trial, a test image is presented, followed by a choice screen
211 showing canonical views of the two possible objects (the object that was not displayed in the test
212 image is referred to as the “distractor” object, but note that objects are equally likely to be
213 distractors and targets). Here, 24 objects were tested, which resulted in 276 binary object
214 discrimination tasks. To neutralize feature attention, these 276 tasks are randomly interleaved
215 (trial by trial), and the global task is referred to as a basic-level, core object recognition task
216 paradigm.

217

218 *Testing human behavior*

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

225

226 Figure 1B illustrates the time course of each behavioral trial, for a particular object
227 discrimination task (zebra versus dog). Each trial initiated with a central black point for 500 ms,
228 followed by 100 ms presentation of a test image containing one foreground object presented
229 under high variation in viewing parameters and overlaid on a random background, as described
230 above (see *Visual images* above). Immediately after extinction of the test image, two choice
231 images, each displaying a single object in a canonical view with no background, were shown to
232 the left and right. One of these two objects was always the same as the object that generated the
233 test image (i.e., the correct object choice), and the location of the correct object (left or right) was
234 randomly chosen on each trial. After clicking on one of the choice images, the subject was
235 queued with another fixation point before the next test image appeared. No feedback was given;
236 human subjects were never explicitly trained on the tasks. Under assumptions of typical
237 computer ergonomics, we estimate that images were presented at 6–8° of visual angle in size,
238 and the choice object images were presented at ± 6 –8° of eccentricity along the horizontal
239 meridian.

240

241 We measured human behavior using the online Amazon MTurk platform (see Figure 1C),
242 which enables efficient collection of large-scale psychophysical data from crowd-sourced
243 “human intelligence tasks” (HITs). The reliability of the online MTurk platform has been
244 validated by comparing results obtained from online and in-lab psychophysical experiments
245 (Majaj, Hong et al. 2015, Rajalingham, Schmidt et al. 2015). We pooled 927,296 trials from
246 1472 human subjects to characterize the aggregate human behavior, which we refer to as the
247 “pooled” human (or “archetypal” human). Each human subject performed only a small number
248 of trials (~xx) on a subset of the images and binary tasks. All 2400 images were used for
249 behavioral testing, but in some of the HITs, we biased the image selection towards the 240
250 primary test images (1424 ± 70 trials/image on this subsampled set, versus 271 ± 93 trials/image on
251 the remaining images, mean \pm SD) to efficiently characterize behavior at image level resolution.
252 Images were randomly drawn such that each human subject was exposed to each image a
253 relatively small number of times (1.5 ± 2.0 trials/image per subject, mean \pm SD), in order to
254 mitigate potential alternative behavioral strategies (e.g. “memorization” of images) that could
255 potentially arise from a finite image set. Behavioral metrics at the object-level (B.O1, B.O2, see
256 Behavioral Metrics) were measured using all 2400 test images, while image-level behavioral
257 metrics (B.I1n, B.I2n) were measured using the 240 primary test images. (We observed
258 qualitatively similar results for those metrics using the full 2400 test images, but we here focus
259 on the primary test images as the larger number of trials leads to lower noise levels).

260

261 Four other human subjects were separately recruited on MTurk to each perform a large
262 number of trials on the same images and tasks ($53,097 \pm 15,278$ trials/subject, mean \pm SD).
263 Behavioral data from these four subjects was not included in the characterization of the pooled
264 human described above, but instead aggregated together to characterize a distinct held-out
265 human pool. This held-out human pool serves to provide a “gold-standard” for benchmarking all
266 other candidate models.

267

268 *Testing monkey behavior*

269 Five adult male rhesus macaque monkeys (*Macaca mulatta*, subjects *M, Z, N, P, B*) were
270 tested on the same basic-level, core object recognition task paradigm described above, with

271 minor modification as described below. All procedures were performed in compliance with
272 National Institutes of Health guidelines and the standards of the Massachusetts Institute of
273 Technology Committee on Animal Care and the American Physiological Society. To efficiently
274 characterize monkey behavior, we used a novel home-cage behavioral system developed in our
275 lab (termed MonkeyTurk, see Fig. 1C). This system leveraged a tablet touchscreen (9" Google
276 Nexus or 10.5" Samsung Galaxy Tab S) and used a web application to wirelessly load the task
277 and collect the data (code available at <https://github.com/dicarloolab/mkturk>). Analogous to the
278 online Amazon Mechanical Turk, which allows for efficient psychophysical assays of a large
279 number (hundreds) of human users in their native environments, MonkeyTurk allowed us to test
280 many monkey subjects simultaneously in their home environment. Each monkey voluntarily
281 initiated trials, and each readily performed the task a few hours each day that the task apparatus
282 was made available to it. At an average rate of ~2,000 trials per day per monkey, we collected a
283 total of 836,117 trials from the five monkey subjects over a period of ~3 months.

284
285 Monkey training is described in detail elsewhere (Rajalingham, Schmidt et al. 2015).
286 Briefly, all monkeys were initially trained on the match-test-image-to-object rule using other
287 images and were also trained on discriminating the particular set of 24 objects tested here using a
288 separate set of training images rendered from these objects, in the same manner as the main
289 testing images. Two of the monkeys subjects (Z and M) were previously trained in the lab
290 setting, and the remaining three subjects were trained using MonkeyTurk directly in their home
291 cages and did not have significant prior lab exposure. Once monkeys reached saturation
292 performance on training images, we began the behavioral testing phase to collect behavior on
293 test images. Monkeys did improve throughout the testing phase, exhibiting an increase in
294 performance between the first and second half of trials of $4\% \pm 0.9\%$ (mean \pm SEM over five
295 monkey subjects). However, the image-level behavioral pattern of the first and second half of
296 trials were highly consistent to each other (B.II consistency of 0.85 ± 0.06 , mean \pm SEM over five
297 monkey subjects), suggesting that monkeys did not significantly alter strategies (e.g. did not
298 “memorize” images) throughout the behavioral testing phase.

299
300 The monkey task paradigm was nearly identical to the human paradigm (see Figure 1B),
301 with the exception that trials were initiated by touching a white “fixation” circle horizontally

302 centered on the bottom third of the screen (to avoid occluding centrally-presented test images
303 with the hand). This triggered a 100ms central presentation of a test image, followed
304 immediately by the presentation of the two choice images (Fig. 1B, location of correct choice
305 randomly assigned on each trial, identical to the human task). Unlike the main human task,
306 monkeys responded by directly touching the screen at the location of one of the two choice
307 images. Touching the choice image corresponding to the object shown in the test image resulted
308 in the delivery of a drop of juice through a tube positioned at mouth height (but not obstructing
309 view), while touching the distractor choice image resulted in a three second timeout. Because
310 gaze direction typically follows the hand during reaching movements, we assumed that the
311 monkeys were looking at the screen during touch interactions with the fixation or choice targets.
312 In both the lab and in the home cage, we maintained total test image size at ~6 degrees of visual
313 angle, and we took advantage of the retina-like display qualities of the tablet by presenting
314 images pixel matched to the display (256 x 256 pixel image displayed using 256 x 256 pixels on
315 the tablet at a distance of 8 inches) to avoid filtering or aliasing effects.

316
317 As with Mechanical Turk testing in humans, MonkeyTurk head-free home-cage testing
318 enables efficient collection of reliable, large-scale psychophysical data but it likely does not yet
319 achieve the level of experimental control that is possible in the head-fixed laboratory setting.
320 However, we note that when subjects were engaged in home-cage testing, they reliably had their
321 mouth on the juice tube and their arm positioned through an armhole. These spatial constraints
322 led to a high level of head position trial-by-trial reproducibility during performance of the task
323 paradigm. Furthermore, when subjects were in this position, they could not see other animals as
324 the behavior box was opaque, and subjects performed the task at a rapid pace 40 trials/minute
325 suggesting that they were not frequently distracted or interrupted. The location of the upcoming
326 test image (but not the location of the object within that test image) was perfectly predictable at
327 the start of each behavioral trial, which likely resulted in a reliable, reproduced gaze direction at
328 the moment that each test image was presented. And the relatively short (but natural and high
329 performing (Cadieu, Hong et al. 2014)) test image duration (100 ms) insured that saccadic eye
330 movements were unlike to influence test image performance (as they generally take ~200 ms to
331 initiate in response to the test image, and thus well after the test image has been extinguished).

332

333 *Testing model behavior*

334 We tested a number of different deep convolutional neural network (DCNN) models on
335 the exact same images and tasks as those presented to humans and monkeys. Importantly, our
336 core object recognition task paradigm is closely analogous to the large-scale ImageNet 1000-way
337 object categorization task for which these networks were optimized and thus expected to perform
338 well. We focused on publicly available DCNN model architectures that have proven highly
339 successful with respect to this benchmark over the past five years: AlexNet (Krizhevsky,
340 Sutskever et al. 2012), NYU (Zeiler and Fergus 2014), VGG (Simonyan and Zisserman 2014),
341 GoogleNet (Szegedy, Zaremba et al. 2013), Resnet (He, Zhang et al. 2016), and Inception-v3
342 (Szegedy, Zaremba et al. 2013). As this is only a subset of possible DCNN models, we refer to
343 these as the $DCNN_{IC}$ (to denote ImageNet-Categorization) visual system model sub-family. For
344 each of the publicly available model architectures, we first used ImageNet-categorization-trained
345 model instances, either using publicly available trained model instances, or training them to
346 saturation on the 1000-way classification task in-house. Training took several days on 1-2 GPUs.
347 The final feature layer of ImageNet trained $DCNN_{IC}$ models corresponds to the probability
348 output of this 1000-way classification task. We adapted these ImageNet-trained models to our
349 24-way object recognition task by re-training the final class probability layer, while holding all
350 other layers fixed. In practice, this was done by extracting features from the penultimate layer of
351 each $DCNN_{IC}$ (i.e. top-most prior to class probability layer), on the same images that were
352 presented to humans and monkeys, and training back-end multi-class logistic regression
353 classifiers to estimate the output class probability for each image. This procedure is illustrated in
354 Figure 1C. To estimate the hit rate of a given image in a given binary classification task, we
355 renormalized the 24-way class probabilities of that image, considering only the two relevant
356 classes, to sum to one. Object-level and image-level behavioral metrics were computed based on
357 these hit rate estimates (as described in *Behavioral Metrics* below).

358

359 From these analyses, we selected the most consistent $DCNN_{IC}$ architecture (Inception-
360 v3), fixed that architecture, and then performed post-hoc analyses in which we varied: the input
361 image sampling, the initial parameter settings prior to training, the filter training images, the type
362 of classifiers used to generate the behavior from the model features, and the classifier training
363 images. To examine input image sampling, we re-trained the Inception-v3 architecture on images

364 from ImageNet that were first spatially filtered to match the spatial sampling of the primate
365 retina (i.e. an approximately exponential decrease in cone density away from the fovea) by
366 effectively simulating a fish-eye transformation on each image. These images were at highest
367 resolution at the “fovea” (i.e. center of the image) with gradual decrease in resolution with
368 increasing eccentricity. To examine the analog of “inter-subject variability”, we constructed
369 multiple trained model instances (“subjects”), where the architecture and training images were
370 held fixed (Inception-v3 and ImageNet, respectively) but the model filter weights initial
371 condition and order of training images were randomly varied for each model instance. To
372 examine the effect of model training, we fine-tuned an ImageNet-trained Inception-v3 model on
373 a synthetic image set consisting of ~6.9 million images of 1049 objects (holding out 50,000
374 images for model validation). These images were generated using the same rendering pipeline as
375 our test images, but the objects were non-overlapping with the 24 test objects presented here. We
376 tested the effect of different classifiers to generate model behavior by testing both multi-class
377 logistic regression and support vector machine classifiers. Additionally, we tested the effect of
378 varying the number of training images used to train those classifiers (20 versus 50 images per
379 class).

380

381 *Behavioral metrics*

382 We measured the object recognition behavior of humans, macaques and DCNN_{IC} models
383 using many test images in 276 interleaved binary object discrimination tasks (see above) To
384 analyze these behavioral data, we here introduce four behavioral (*B*) metrics of increasing
385 richness, but requiring increasing amounts of data to measure reliably. Each behavioral metric
386 computes a pattern of unbiased behavioral performance, using a sensitivity index: $d' =$
387 $Z(\text{HitRate}) - Z(\text{FalseAlarmRate})$, where Z is the inverse of the cumulative Gaussian
388 distribution. The various metrics differ in the resolution at which hit rates and false alarm rates
389 are computed. Table 1 summarizes four behavioral metrics, varying the hit-rate resolution
390 (image-level or object-level) and the false-alarm resolution (one-versus-all or one-versus-other).
391 Briefly, the one-versus-all object-level performance metric (termed B.O1) estimates the
392 discriminability of each object from all other objects, pooling across all distractor object choices.
393 Since we here tested 24 objects, the B.O1 metric measured here has 24 independent values. The
394 one-versus-other object-level performance metric (termed B.O2) estimates the discriminability of

395 each specific pair of objects, or the pattern of pairwise object confusions. Since we here tested
396 276 interleaved binary object discrimination tasks, the B.O2 metric measure here has 276
397 independent values (the off-diagonal elements on one half of the 24x24 symmetric matrix). The
398 one-versus-all image-level performance metric (termed B.I1) estimates the discriminability of
399 each image from all other objects, pooling across all 23 possible distractor choices. Since we
400 here focused on the primary image test set of 240 images (10 per object, see above), the B.I1
401 metric measured here has 240 independent values. Finally, the one-versus-other image-level
402 performance metric (termed B.I2) estimates the discriminability of each image from each
403 distractor object. Since we here focused on the primary image test set of 240 images (10 per
404 object, see above) with 23 distractors, the B.I1 metric measured here has 5520 independent
405 values.

406

407 Naturally, object-level and image-level behavioral patterns are tightly linked. For
408 example, images of a particularly difficult-to-discriminate object would inherit lower
409 performance values on average as compared to images from a less difficult-to-discriminate
410 object. To isolate the behavioral variance that is specifically driven by image variation and not
411 simply predicted by the objects (and thus already captured by the B.O1 and B.O2 metrics), we
412 estimated normalized image-level behavioral metrics by subtracting the mean performance
413 values over all images of the same object and task. This process is schematically illustrated in
414 Figure 3A. We focus on these normalized image-level behavioral metrics (termed B.I1n, B.I2n)
415 for image-level comparisons between models and primates (see Results).

416

417 *Behavioral Consistency*

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

426 Pearson correlation over all the independent measurements in the metric from the model (\mathbf{m}) and
427 the human (\mathbf{h}), and we then normalize that raw Pearson correlation by the geometric mean of the
428 split-half internal reliability of the same behavioral metric measured for each system: $\tilde{\rho}(\mathbf{m}, \mathbf{h}) =$
429 $\frac{\rho(\mathbf{m}, \mathbf{h})}{\sqrt{\rho(\mathbf{m}, \mathbf{m})\rho(\mathbf{h}, \mathbf{h})}}$.

430
431 Since all correlations in the numerator and denominator were computed using the same
432 amount of trial data (exactly half of the trial data), we did not need to make use of any prediction
433 formulas (e.g. extrapolation to larger number of trials using Spearman-Brown prediction
434 formula). This procedure was repeated 10 times with different random split-halves of trials. Our
435 rationale for using a noise-adjusted correlation measure for consistency was to account for
436 variance in the behavioral patterns that arises from “noise,” i.e., variability that is not replicable
437 by the experimental condition (image and task) and thus that no model can be expected to predict
438 (Johnson, Hsiao et al. 2002).

439
440 *Characterization of Residuals*

441 In addition to measuring the similarity between the behavioral patterns of primates and
442 models (using consistency analyses, as described above), we examined the corresponding
443 differences, or “residual behavioral patterns.” Each candidate visual system model’s residual
444 behavioral pattern was estimated as the residual of a linear least squares regression on the human
445 pool data (one behavioral performance value per test image, thus 240 values) and we included a
446 free intercept parameter. This procedure effectively captures the differences between human and
447 model behavior after accounting for overall performance differences. Residual patterns were
448 estimated on disjoint split-halves of trials, repeating 10 times with random trial permutations. We
449 focused on the normalized one-versus-all image-level performance pattern (B.I1n) to reliably
450 measure image-level differences between primates and models as that metric showed a clear
451 difference between DCNN_{IC} models and primates, and the behavioral residual can be interpreted
452 based only the test images (i.e. we can assign a residual per image).

453
454 To examine the extent to which the difference between each model and humans is
455 reliably shared across different models, we measured the Pearson correlation between the
456 residual patterns of pairs of models. Residual similarity was quantified as the proportion of

457 shared variance, defined as the square of the noise-adjusted correlation between residual patterns
458 (the noise-adjustment was done as defined in equation above). Correlations of residual patterns
459 were always computed across distinct split-halves of data, to avoid introducing spurious
460 correlations from subtracting common noise in the human data. We measured the residual
461 consistency between all pairs of tested models, holding both architecture and optimization
462 procedure fixed (between instances of the ImageNet-categorization trained Inception-v3 model,
463 varying in filter initial conditions), varying the architecture while holding the optimization
464 procedure fixed (between all tested ImageNet-categorization trained DCNN architectures), and
465 holding the architecture fixed while varying the optimization procedure (between ImageNet-
466 categorization trained Inception-v3 and synthetic-categorization fine-tuned Inception-v3
467 models). This analysis addresses not only the reliability of the failure of $DCNN_{IC}$ models to
468 predict human behavior (deviations from humans), but also the relative importance of the
469 characteristics defining similarities within the model sub-family (namely, the architecture and the
470 optimization procedure). We first performed this analysis for behavioral patterns over the 240
471 primary test images, and subsequently zoomed in on subsets of images that humans found to be
472 particularly difficult. This image selection was made relative to the distribution of image-level
473 performance of held-out human subjects (B.I1 metric from four subjects); difficult images were
474 defined as ones with performance below the 50th and 25th percentiles of this distribution.

475
476 To examine whether the difference between each model and humans can be explained by
477 simple human-interpretable stimulus attributes, we regressed each $DCNN_{IC}$ model's residual
478 pattern from image attributes, including viewpoint attributes (e.g. object size, eccentricity, pose)
479 and pixel attributes (e.g. mean luminance, background spatial frequency, segmentation-index).
480 Briefly, we constructed a design matrix from the image attributes (using individual attributes,
481 groups of attributes, or all attributes), and used multiple linear least squares regression to predict
482 the image-level residual pattern. The multiple linear regression was tested using two-fold cross-
483 validation over trials. The relative importance of each attribute (or groups of attributes) was
484 quantified using the proportion of explainable variance (i.e. variance remaining after accounting
485 for noise variance) explained from the residual pattern.

486
487 *Primate zone*

488 In this work, we are primarily concerned with the behavior of an “archetypal human”,
489 rather than the behavior of any given individual human subject. We operationally defined this
490 concept as the common behavior over many humans, obtained by pooling together trials from a
491 large number of individual human subjects and treating this human pool as if it were acquired
492 from a single behaving agent. Due to inter-subject variability, we do not expect any given
493 human or monkey subject to be perfectly consistent (i.e. have consistency of 1.0) with this
494 archetypal human. Given current limitations of monkey psychophysics, we are not yet able to
495 measure the behavior of very large number of monkey subjects at high resolution and
496 consequently cannot directly estimate the consistency of the corresponding “archetypal monkey”
497 to the human pool. Rather, we indirectly estimated this consistency by first measuring
498 consistency as a function of number of individual subjects pooled together (n), and extrapolating
499 the consistency estimate for pools of very large number of subjects (as n approaches infinity).
500 Extrapolations were done using least squares fitting of an exponential function $\tilde{\rho}(n) = a + b \cdot$
501 e^{-cn} (see Figure 4).

502
503 For each behavioral metric, we defined a “primate zone” as the range of consistency
504 values delimited by consistency estimates $\tilde{\rho}_{M\infty}$ and $\tilde{\rho}_{H\infty}$ as lower and upper bounds respectively.
505 $\tilde{\rho}_{M\infty}$ corresponds to the extrapolated estimate of consistency relative to the human pool of a
506 large (i.e. infinitely many subjects) pool of rhesus macaque monkeys; $\tilde{\rho}_{H\infty}$ is by definition equal
507 to 1.0. Thus, the primate zone defines a range of consistency values that correspond to models
508 that accurately capture the behavior of the human pool, at least as well as an extrapolation of our
509 monkey sample. In this work, we defined this range of behavioral consistency values as the
510 criterion for success for computational models of primate visual object recognition behavior.

511
512 To make a global statistical inference about whether models sampled from the $DCNN_{IC}$
513 sub-family meet or fall short of this criterion for success, we attempted to reject the hypothesis
514 that, for a given behavioral metric, the human consistency of $DCNN_{IC}$ models is within the
515 primate zone. To test this hypothesis, we estimate the empirical probability that the distribution
516 of human consistency values, estimated over different model instances within this family, could
517 produce human consistency values within the primate zone. Specifically, we estimated a p-value
518 for each behavioral metric using the following procedure: We first estimated an empirical

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

529

530 **RESULTS**

531

532 In the present work, we systematically compared the basic level core object recognition
533 behavior of primates and state-of-the-art artificial neural network models using a series of
534 behavioral metrics (B) ranging from low to high resolution within a two-alternative forced
535 choice match-to-sample paradigm. The behavior of each visual system, whether biological or
536 computational, was tested on the same 2400 images (24 objects, 100 images/object) in the same
537 276 interleaved binary object recognition tasks. Each system's behavior was characterized at
538 multiple resolutions (see *Behavioral metrics* in Methods) and directly compared to the
539 corresponding behavioral metric of the archetypal human (defined as the average behavior over a
540 large pool of human subjects tested; see Methods). The overarching logic of this study is that, if
541 two visual systems are equivalent, they should produce statistically indistinguishable behavioral
542 metrics (B).

543

544 *Object-level behavioral comparison*

545 We first examined the pattern of one-versus-all object-level behavior (termed “B.O1
546 metric”) computed across all images and possible distractors. Since we tested 24 objects here,
547 the B.O1 metric vector is 24 dimensional. Figure 2A shows the B.O1 metric vector for the
548 pooled human (pooling $n=1472$ human subjects), pooled monkey (pooling $n=5$ monkey
549 subjects), and several DCNN_{IC} models as 24-dimensional vectors using a color scale. Each bin

550 corresponds to the system's discriminability of one object against all others that were tested (i.e.
551 all other 23 objects). The color scales span each pattern's full performance range, and warm
552 colors indicate lower discriminability. For example, red indicates that the tested visual system
553 found the object corresponding to that element of the vector to be very challenging to
554 discriminate from other objects (on average over all 23 discrimination tests, and on average over
555 all images). Figure 2B directly compares the B.O1 metric vector computed from the behavioral
556 output of two visual system models—a pixel model (top panel) and a DCNN_{IC} model (Inception-
557 v3, bottom panel)—against that of the human BO1 metric vector. We observe a tighter
558 correspondence to the human behavioral pattern for the DCNN_{IC} model visual system than for
559 the baseline pixel model visual systems. We quantified that similarity using a noise-adjusted
560 correlation between each pair of B.O1 vectors (termed *consistency*, following (Johnson, Hsiao et
561 al. 2002)); the noise adjustment means that a visual system that is identical to the human pool
562 will have an expected human consistency score of 1.0, even if it has irreducible trial-by-trial
563 stochasticity; see Methods). Figure 2C shows the B.O1 human consistency for each of the tested
564 model visual systems. We additionally tested the behavior of a held-out pool of four human
565 subjects (black dot) and a pool of five macaque monkey subjects (gray dot), and we observed
566 that both yielded B.O1 vectors that were highly consistent to the human pool ($\tilde{\rho} = 0.90, 0.97$ for
567 monkey pool and held-out human pool, respectively). We defined a range of consistency values,
568 termed the “primate zone” (shaded gray area), delimited by extrapolated human consistency
569 estimates of large pools of macaques and humans (see Methods, Figure 4). With respect to the
570 B.O1 metric, all tested DCNN_{IC} visual system models were either within or very close to this
571 zone, while the baseline pixel visual system model and the low-level V1 visual system model
572 were not ($\tilde{\rho} = 0.40, 0.67$ for pixels and V1 models, respectively). Based on the B.O1 behavioral
573 metric alone, the hypothesis that the human consistency of DCNN_{IC} models is within the primate
574 zone could not be rejected ($p = 0.54$, exact test, see Methods).

575

576 Next, we compared the behavior of the visual systems at a slightly higher level of
577 resolution. Specifically, instead of pooling over all discrimination tasks for each object, we
578 computed the mean discriminability of each of the 276 pairwise discrimination tasks (still
579 pooling over images within each of those tasks). This yields a symmetric matrix that is referred
580 to here as the B.O2 metric. Figure 2D shows the B.O2 metric for pooled human, pooled monkey,

581 and several DCNN_{IC} visual system models as 24x24 symmetric matrices. Each bin (i,j)
582 corresponds to the system's discriminability of objects i and j , where warmer colors indicate
583 lower performance; color scales are not shown but span each pattern's full range. We observed
584 strong qualitative similarities between the pairwise object confusion patterns of all of the high
585 level visual systems (e.g. camel and dog are often confused with each other by all three systems).
586 This similarity is quantified in Figure 2E, which shows the consistency relative to the human
587 pool of all examined visual system models with respect to this metric. Similar to the B.O1
588 metric, we observed that both a pool of macaque monkeys and a held-out pool of humans are
589 highly consistent to the human pool with respect to this metric ($\tilde{\rho} = 0.77, 0.94$ for monkeys,
590 humans respectively). Also similar to the B.O1 metric, we found that all DCNN_{IC} visual system
591 models are highly consistent with the human pool ($\tilde{\rho} > 0.8$) while the baseline pixel visual
592 system model and the low-level V1 visual system model were not ($\tilde{\rho} = 0.41, 0.57$ for pixels, V1
593 models respectively). Indeed, all DCNN_{IC} visual system models are within the defined "primate
594 zone" of human consistency. Again, based on the B.O2 behavioral metric, the hypothesis that the
595 human consistency of the DCNN_{IC} models is within the primate zone could not be rejected ($p =$
596 0.99 , exact test).

597

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

609

610 *Image-level behavioral comparison*

611 To isolate purely image-level behavioral variance, i.e. variance that is not predicted by
612 the object and thus already captured by the B.O1 metric, we focused our analyses on normalized
613 image-level performance patterns. This normalization procedure is schematically illustrated in
614 Figure 3A for the one-versus-all image-level performance pattern (240-dimensional, 10
615 images/object) to obtain the normalized one-versus-all image-level behavioral metric (termed
616 B.I1n metric, see Methods). Figure 3B shows the B.I1n metric for the pooled human, pooled
617 monkey, and several DCNN_{IC} models as 240 dimensional vectors. Each bin's color corresponds
618 to the discriminability of a single image against all distractor options (after subtraction of object-
619 level discriminability, see Figure 3A), where warmer colors indicate lower values; color scales
620 are not shown but span each pattern's full range. Figure 3D shows the consistency to the human
621 pool with respect to the B.I1n metric for all tested models. Unlike with object-level behavioral
622 metrics, we now observe a divergence between DCNN_{IC} models and primates. Both the monkey
623 pool and the held-out human pool remain highly consistent with the pooled human with respect
624 to this metric ($\tilde{\rho} = 0.77, 0.96$ for monkeys, humans respectively), but all DCNN_{IC} models were
625 significantly less consistent (Inception-v3: $\tilde{\rho} = 0.62$) and well outside of the defined "primate
626 zone" of I1_c consistency to the human pool. Indeed, based on the B.I1n behavioral metric, the
627 hypothesis that the human consistency of DCNN_{IC} models is within the primate zone is strongly
628 rejected ($p = 6.16e-8$, exact test, see Methods).

629
630 We can zoom in further on this metric by examining not only the overall performance for
631 a given image but also the object confusions for each image, i.e. the additional behavioral
632 variation that is due not only to the test image but to the interaction of that test image with the
633 alternative (incorrect) object choice that is provided after the test image (see Fig. 1B). This is the
634 highest level of behavioral accuracy resolution that our task design allows. In raw form, it
635 corresponds to one-versus-other image-level confusion matrix, where the size of that matrix is
636 the total number of images by the total number of objects (here, 240x24). Each bin (i,j)
637 corresponds to the behavioral discriminability of a single image i against distractor object j .
638 Again, we isolate variance that is not predicted by object-level performance by subtracting the
639 average performance on this binary task (mean over all images) to convert the raw matrix B.I2
640 above into the normalized matrix, referred to as B.I2n. Figure 3D shows the B.I2n metric as
641 240x24 matrices for the pooled human, pooled monkey and top DCNN_{IC} visual system models.

642 Color scales are not shown but span each pattern’s full range; warmer colors correspond to
643 images with lower performance in a given binary task, relative to all images of that object in the
644 same task. Figure 3E shows the human consistency with respect to the B.I2n metric for all tested
645 visual system models. Extending our observations using the vector of image difficulties (B.I1n),
646 we observe a similar divergence between primates and DCNN_{IC} visual system models on the
647 matrix pattern of image-by-distractor difficulties (I2n). Specifically, both the monkey pool and
648 held-out human pool remain highly consistent with the pooled human ($\tilde{\rho} = 0.75, 0.77$ for
649 monkeys, humans respectively), while all tested DCNN_{IC} models are significantly less consistent
650 (Inception-v3: $\tilde{\rho} = 0.53$) falling well outside of the defined “primate zone” of I2n consistency to
651 the human pool. Once again, based on the B.I2n behavioral metric, the hypothesis that the human
652 consistency of DCNN_{IC} models is within the primate zone is strongly rejected ($p = 3.17e-18$,
653 exact test, see Methods).

654

655 *Natural subject-to-subject variation*

656 For each behavioral metric (B.O1, BO2, B.I1n, BI2n), we defined a “primate zone” as the
657 range of consistency values delimited by consistency estimates $\tilde{\rho}_{M\infty}$ and $\tilde{\rho}_{H\infty}$ as lower and upper
658 bounds respectively. $\tilde{\rho}_{M\infty}$ corresponds to the extrapolated estimate of the human (pool)
659 consistency of a large (i.e. infinitely many subjects) pool of rhesus macaque monkeys. Thus, the
660 fact that a particular tested visual system model falls outside of the primate zone can be
661 interpreted as a failure of that visual system model to accurately predict the behavior of the
662 archetypal human at least as well as the archetypal monkey.

663

664 However, from the above analyses, it is not yet clear whether a visual system model that
665 fails to predict the archetypal human might nonetheless accurately correspond to one or more
666 individual human subjects found within the natural variation of the human population. Given the
667 difficulty of measuring individual subject behavior at the resolution of single images for large
668 numbers of human and monkey subjects, we could not yet directly test this hypothesis. Instead,
669 we examined it indirectly by asking whether an archetypal model—that is a pool that includes an
670 increasing number of model “subjects”—would approach the human pool. We simulated model
671 inter-subject variability by retraining a fixed DCNN architecture with a fixed training image set
672 with random variation in the initial conditions and order of training images. This procedure

673 results in models that can still perform the task but with slightly different learned weight values.
674 We note that this procedure is only one possible choice of generating inter-subject variability
675 within each visual system model type, a choice that is an important open research direction that
676 we do not address here. From this procedure, we constructed multiple trained model instances
677 (“subjects”) for a fixed DCNN architecture, and asked whether an increasingly large pool of
678 model “subjects” better captures the behavior of the human pool, at least as well as a monkey
679 pool. This post-hoc analysis was conducted for the most human consistent DCNN architecture
680 (Inception-v3).

681

682 Figure 4A shows the measured human consistency for each of the four behavioral
683 metrics, for subject pools of varying size (number of subjects n) of rhesus macaque monkeys
684 (black) and ImageNet-trained Inception-v3 models (blue). The human consistency increases with
685 growing number of subjects for both visual systems across all behavioral metrics. To estimate
686 the expected human consistency for a pool of infinitely many monkey or model subjects, we fit
687 an exponential function mapping n to the mean consistency values and obtained a parameter
688 estimate for the asymptotic value (see Methods). We note that estimated asymptotic values are
689 not significantly beyond the range of the measured data—the human consistency of a pool of
690 five monkey subjects reaches within 97% of the human consistency of an estimated infinite pool
691 of monkeys for all metrics—giving credence to the extrapolated consistency values. This
692 analysis suggests that under this model of inter-subject variability, a pool of Inception-v3
693 subjects accurately capture archetypal human behavior at the resolution of objects (B.O1, B.O2)
694 by our primate zone criterion (see Figure 4A, first two panels). In contrast, even a large pool of
695 Inception-v3 subjects still fails at its final asymptote to accurately capture human behavior at the
696 image-level (B.I1n, B.I2n) (Figure 4A, last two panels).

697

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

699 Next, we wondered if some relatively simple changes to the DCNN_{IC} visual system
700 models tested here could bring them into better correspondence with the primate visual system
701 behavior (with respect to B.I1n and B.I2n metrics). Specifically, we considered and tested the
702 following modifications to the DCNN_{IC} model visual system that scored the highest in our
703 benchmarks (Inception-v3): we (1) changed the input to the model to be more primate-like in its

704 retinal sampling (Inception-v3 + retina-like), (2) changed the transformation (aka “decoder”)
705 from the internal model feature representation into the behavioral output by augmenting the
706 number of decoder training images or changing the decoder type (Inception-v3 + SVM,
707 Inception-v3 + classifier_train), and (3) modified all of the internal filter weights of the model
708 (aka “fine tuning”) by augmenting its ImageNet training with additional images drawn from the
709 same distribution as our test images (Inception-v3 + synthetic-fine-tune). While some of these
710 modifications (e.g. fine-tuning on synthetic images and increasing the number of classifier
711 training images) had the expected effect of increasing mean overall performance (not shown), we
712 found that none of these modifications led to a significant improvement in its human consistency
713 on the behavioral metrics (Figure 4B). Thus, the failure of current DCNN_{IC} models to accurately
714 capture the image-level behavioral patterns of primates cannot be rescued by simple
715 modifications on a fixed architecture.

716

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

718 Taken together, Figures 2, 3 and 4 suggest that current DCNN_{IC} visual system models fail
719 to accurately capture the image-level behavioral patterns of humans and monkeys. To further
720 examine this failure in the hopes of providing clues for model improvement, we examined the
721 residual image-level behavioral patterns of all the visual system models, relative to the pooled
722 human. For each model, we computed its residual image-level behavioral pattern as the
723 difference (positive or negative) of a linear least squares regression of the model predictions with
724 the human pool observations. For this analysis, we focused on the B.I1n metric as it showed a
725 clear divergence of DCNN_{IC} models and primates, and the behavioral residual can be interpreted
726 based only on the test images (whereas B.I2n depends on the interaction between test images and
727 distractor choice). We first asked to what extent the residual image-level behavioral patterns are
728 shared between different visual system models.

729

730 Figure 5A shows the similarity between the residual image-level patterns of all pairs of
731 models; the color of bin (i,j) indicates the proportion of explainable variance that is shared
732 between the residual image-level patterns of visual systems i and j . For ease of interpretation, we
733 ordered visual system models based on their architecture and optimization procedure and
734 partitioned this matrix into four distinct regions. Each region compares the residuals of a

735 “source” model group with fixed architecture and optimization procedure (five Inception-v3
736 models optimized for categorization on ImageNet, varying only in initial conditions and training
737 image order) to a “target” model group. The target groups of models for each of the four regions
738 are: 1) the pooled monkey, 2) other DCNN_{IC} models from the source group, 3) DCNN_{IC} models
739 that differ in architecture but share the optimization procedure of the source group models and 4)
740 DCNN_{IC} models that differ slightly using an augmented optimization procedure but share the
741 architecture of the source group models. Figure 5B shows the mean (\pm SD) variance shared in the
742 residuals averaged within these four regions for all images (black dots), as well as for images
743 that humans found to be particularly difficult (blue and red dots, selected based on held-out
744 human data, see Methods). First, consistent with the results shown in Figure 3, we note that the
745 residual image-level patterns of this particular DCNN_{IC} model are not well shared with the
746 pooled monkey ($r^2=0.39$ in region 1), and this phenomenon is more pronounced for the images
747 that humans found most difficult ($r^2=0.17$ in region 1). However, this relatively low correlation
748 between model and primate residuals is not indicative of spurious model residuals, as the image-
749 level residual patterns were highly reliable between different instances of this fixed DCNN_{IC}
750 model, across random training initializations (region 2: $r^2=0.79$, 0.77 for all and most difficult
751 images, respectively). Interestingly, residual patterns were still largely shared with other DCNN_{IC}
752 models with vastly different architectures (region 3: $r^2=0.70$, 0.65 for all and most difficult
753 images, respectively). However, residual patterns were more strongly altered when the visual
754 training diet of the same architecture was altered (region 4: $r^2=0.57$, 0.46 for all and most
755 difficult images respectively, cf. region 3). Taken together, these results indicate that the images
756 where DCNN_{IC} visual system models diverged from humans (and monkeys) were not spurious
757 but were rather highly reliable across different model architectures, demonstrating that current
758 DCNN_{IC} models systematically and similarly diverge from primates.

759

760 To look for clues for model improvement, we asked what, if any, characteristics of
761 images might account for this divergence of models and primates. We regressed the residual
762 image-level behavioral pattern of the Inception-v3 architecture on a range of image attributes.
763 Specifically, we considered both object viewpoint attributes (the size, eccentricity, and pose of
764 the object) and pixel attributes (mean luminance, background spatial frequency, segmentation
765 index) of each image. We used multivariate regressions to predict the residual pattern from

766 groups of several image attributes (e.g. from all attributes), and also considered each attribute
767 individually using univariate regressions. Figure 6A shows example images (sampled from the
768 full set of 2400 images) with increasing attribute value for each of these six image attributes.
769 While the DCNN_{IC} models were not directly optimized to display primate-like performance
770 dependence on such attributes, we observed that the Inception-v3 visual system model
771 nonetheless exhibited qualitatively similar performance dependencies as primates (see Figure
772 6B). For example, humans (black), monkeys (gray) and the Inception-v3 model (blue) all
773 performed better, on average, for images in which the object is in the center of gaze (low
774 eccentricity) and large in size. The similarity of the patterns in Figure 6B between primates and
775 the DCNN_{IC} visual system models is not perfect but is striking, particularly in light of the fact
776 that these models were not optimized to produce these patterns. However, this similarity is
777 analogous to the similarity in the B.O1 and B.O2 metrics in that it only holds on average over
778 many images. Looking more closely at the image-by-image comparison, we again found that the
779 DCNN_{IC} models failed to capture a large portion of the image-by-image variation (Figure 3). In
780 particular, Figure 6C shows the proportion of variance explained by specific image attributes for
781 the residual, patterns of monkeys (dark gray), Inception-v3 models (dark blue), and all DCNN_{IC}
782 models (light blue). We found that, taken together, all six of these image attributes explained
783 only ~10% of the variance in the image-wise residual between humans and DCNN_{IC}.
784 Furthermore, we found that pixel attributes, rather than viewpoint attributes, contributed the
785 majority of this explanatory power. Each individual attribute could explain at most a small
786 amount of residual behavioral variance (<5% of the explainable variance). In sum, these analyses
787 show that some behavioral effects that might provide intuitive clues to modify the DCNN_{IC}
788 models are already in place in those models (e.g. a dependence on eccentricity). But the
789 quantitative image-by-image analyses of the remaining unexplained variance (Figure 6C) argue
790 that the DCNN_{IC} visual system models' failure to capture primate image-level performance
791 patterns cannot be further accounted for by these simple image attributes and likely stem from
792 other factors.

793

794 **DISCUSSION**

795

796 Broadly, our scientific goal is to discover computational models that quantitatively
797 explain the neuronal mechanisms underlying primate invariant object recognition behavior. To
798 this end, previous work had shown that specific artificial neural network models, drawn from a
799 large family of deep convolutional neural networks and optimized to achieve high levels of
800 object categorization performance on large-scale image-sets, accurately capture the coarse
801 behavioral patterns of primates in core object recognition tasks while the internal hidden neurons
802 of those same models also predict a large fraction of primate ventral stream neural response
803 variance to images (Cadieu, Hong et al. 2014, Khaligh-Razavi and Kriegeskorte 2014, Yamins,
804 Hong et al. 2014, Güçlü and van Gerven 2015, Rajalingham, Schmidt et al. 2015, Kheradpisheh,
805 Ghodrati et al. 2016, Kubilius, Bracci et al. 2016). For clarity, we here referred to this sub-family
806 of models as DCNN_{IC} (to denote ImageNet-Categorization training), so as to distinguish them
807 from all possible models in the DCNN family, and more broadly, from the super-family of all
808 ANNs. In this work, we directly compared leading DCNN_{IC} models to primates (humans and
809 monkeys) with respect to their behavioral patterns at both object and image level resolution in
810 the domain of core object recognition. Our primary novel result is that leading DCNN_{IC} models
811 fail to fully replicate the image-level behavioral patterns of primates. An important related claim
812 is that rhesus monkeys are more consistent with the archetypal human than any of the tested
813 DCNN_{IC} models.

814
815 While it had previously been shown that DCNN_{IC} models can diverge from human
816 behavior on specifically chosen adversarial images (Szegedy, Zaremba et al. 2013), a strength of
817 our work is that we did not optimize images to induce failure but instead randomly sampled the
818 image generative parameter space broadly. Furthermore, we showed that the failure of current
819 DCNN_{IC} models to accurately predict primate behavioral patterns cannot be explained by simple
820 image attributes (e.g. object viewpoint meta-parameters and low-level image statistics) and
821 cannot be rescued by simple model modifications (input image sampling, model training, and
822 classifier variations). Taken together, these results expose a general failure of current DCNN_{IC} to
823 fully replicate the image-level behavioral patterns of primates and suggest that new ANN models
824 are needed to more precisely capture the neural mechanisms underlying primate object vision.

825

826 With regards to new ANN models, we can attempt to make prospective inferences about
827 new and untested models from the data presented here. Based on the observed distribution of
828 image-level behavioral consistency values for the tested DCNN_{IC} models, one could infer that yet
829 untested model instances sampled identically (i.e. from the same model sub-family) are highly
830 likely to have similarly inadequate image-level behavioral consistency with primates. While we
831 cannot rule out the possibility that at least one model instance within the DCNN_{IC} class fully
832 matches image-level human patterns, the probability of sampling such a model is vanishingly
833 small ($p < 10^{-18}$ for B.I2n consistency, estimated using exact test using Gaussian kernel density
834 estimation, see Methods, Results). An important caveat of this inference is that we may have
835 poorly estimated the consistency distribution, as we did not exhaustively sample this model
836 family. In particular, if the model sampling process is non-stationary over time (e.g. increases in
837 computational power over time allows larger models to be successfully trained), the consistency
838 of new (yet to be sampled) models may lie outside the currently estimated distribution.
839 Consistent with the latter, we observed that current DCNN_{IC} cluster into two distinct
840 “generations” separated in time (before/after the year 2015; e.g. Inception-v3 improves over
841 Alexnet though both lie outside the primate zone in Figure 3). Thus, following this trend, it is
842 possible that the evolution of “next-generation” models within the DCNN_{IC} sub-family could
843 meet the criterion for success of primate-like behavior.

844
845 Alternatively, it is possible that new DCNN_{IC} models would also fail to capture primate-
846 like image-level behavior, suggesting that either the architectural limitations (e.g. convolutional,
847 feed-forward) and/or the optimization procedure (including the diet of visual images) that define
848 this model sub-family are fundamentally limiting. Thus, ANN model sub-families utilizing
849 different architectures (e.g. recurrent neural networks) and/or optimized for different behavioral
850 goals (e.g. loss functions other than object classification performance, and/or images other than
851 category-labeled ImageNet images) may be necessary to accurately capture primate behavior. To
852 this end, we propose that testing individual changes to the DCNN_{IC} models—each creating a new
853 ANN model sub-family—may be the best way forward, as DCNN_{IC} models currently best
854 explain both the behavioral and neural phenomena of core object recognition.

855

856 To reach that goal of finding a new ANN model sub-family that is an even better
857 mechanistic model of the primate ventral visual stream, we propose that even larger-scale, high-
858 resolution behavioral measurements than previously used, such as expanded versions of the
859 patterns of image-level performance presented here, could serve as a useful top-down
860 optimization constraint. Not only do these high-resolution behavioral metrics have the statistical
861 power to reject the currently leading ANN models, but they can also be efficiently collected at
862 very large scale, in contrast to other constraint data (e.g. large-scale neuronal measurements).
863 Indeed, current technological tools for high-throughput psychophysics in humans and monkeys
864 (e.g. Amazon Mechanical Turk for humans, Monkey Turk for rhesus monkeys) enable time- and
865 cost-efficient collection of large-scale behavioral datasets, such as the ~1 million behavioral
866 trials obtained for the current work. These systems trade off an increase in efficiency with a
867 decrease in experimental control. For example, we did not impose experimental constraints on
868 subjects' acuity and we can only infer likely head and gaze position. Previous work has shown
869 that patterns of behavioral performance on object recognition tasks from in-lab and online
870 subjects were equally reliable and virtually identical (Majaj, Hong et al. 2015), but it is not yet
871 clear to what extent this holds at the resolution of individual images, as one might expect that
872 variance in performance across images is more sensitive to precise head and gaze location. For
873 this reason, we refrain from making strong inferences from small behavioral differences, such as
874 the difference between humans and monkeys. Nevertheless, we argue that this sacrifice in exact
875 experimental control while retaining sufficient power for model comparison is a good tradeoff
876 for the large-scale, high-resolution behavioral datasets that could be efficiently collected in both
877 humans and monkeys, specifically toward the goal of constraining future models of the primate
878 ventral visual stream.
879

880 **REFERENCES**

- 881
882 Cadieu, C. F., et al. (2014). "Deep neural networks rival the representation of primate IT cortex
883 for core visual object recognition." PLoS computational biology **10**(12): e1003963.
- 884
885 DiCarlo, J. J. and D. D. Cox (2007). "Untangling invariant object recognition." Trends in
886 cognitive sciences **11**(8): 333-341.
- 887
888 DiCarlo, J. J., et al. (2012). "How does the brain solve visual object recognition?" Neuron **73**(3):
889 415-434.
- 890
891 Dodge, S. and L. Karam (2017). "A Study and Comparison of Human and Deep Learning
892 Recognition Performance Under Visual Distortions." arXiv preprint arXiv:1705.02498.
- 893
894 Geirhos, R., et al. (2017). "Comparing deep neural networks against humans: object recognition
895 when the signal gets weaker." arXiv preprint arXiv:1706.06969.
- 896
897 Goodfellow, I. J., et al. (2014). "Explaining and harnessing adversarial examples." arXiv preprint
898 arXiv:1412.6572.
- 899
900 Güçlü, U. and M. A. van Gerven (2015). "Deep neural networks reveal a gradient in the
901 complexity of neural representations across the ventral stream." Journal of Neuroscience
902 **35**(27): 10005-10014.
- 903
904 He, K., et al. (2016). Deep residual learning for image recognition. Proceedings of the IEEE
905 conference on computer vision and pattern recognition.
- 906
907 Hosseini, H., et al. (2017). "On the Limitation of Convolutional Neural Networks in Recognizing
908 Negative Images." human performance **4**(5): 6.
- 909
910 Huynh, D. Q. (2009). "Metrics for 3D rotations: Comparison and analysis." Journal of
911 Mathematical Imaging and Vision **35**(2): 155-164.
- 912
913 Johnson, K. O., et al. (2002). "Neural coding and the basic law of psychophysics." The
914 Neuroscientist **8**(2): 111-121.
- 915
916 Khaligh-Razavi, S.-M. and N. Kriegeskorte (2014). "Deep supervised, but not unsupervised,
917 models may explain IT cortical representation." PLoS computational biology **10**(11): e1003915.
- 918
919 Kheradpisheh, S. R., et al. (2016). "Deep networks can resemble human feed-forward vision in
920 invariant object recognition." Scientific reports **6**: 32672.

- 921
922 Krizhevsky, A., et al. (2012). Imagenet classification with deep convolutional neural networks.
923 Advances in neural information processing systems.
- 924
925 Kubilius, J., et al. (2016). "Deep neural networks as a computational model for human shape
926 sensitivity." PLoS computational biology **12**(4): e1004896.
- 927
928 LeCun, Y., et al. (2015). "Deep learning." Nature **521**(7553): 436-444.
- 929
930 Majaj, N. J., et al. (2015). "Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing
931 Rates Accurately Predict Human Core Object Recognition Performance." The Journal of
932 Neuroscience **35**(39): 13402-13418.
- 933
934 Nguyen, A., et al. (2015). Deep neural networks are easily fooled: High confidence predictions
935 for unrecognizable images. Proceedings of the IEEE Conference on Computer Vision and
936 Pattern Recognition.
- 937
938 Pinto, N., et al. (2008). "Why is real-world visual object recognition hard?" PLoS computational
939 biology **4**(1): e27.
- 940
941 Rajalingham, R., et al. (2015). "Comparison of Object Recognition Behavior in Human and
942 Monkey." The Journal of Neuroscience **35**(35): 12127-12136.
- 943
944 Rosch, E., et al. (1976). "Basic objects in natural categories." Cognitive psychology **8**(3): 382-
945 439.
- 946
947 Simonyan, K. and A. Zisserman (2014). "Very deep convolutional networks for large-scale
948 image recognition." arXiv preprint arXiv:1409.1556.
- 949
950 Szegedy, C., et al. (2013). "Intriguing properties of neural networks." arXiv preprint
951 arXiv:1312.6199.
- 952
953 Ullman, S. and G. W. Humphreys (1996). High-level vision: Object recognition and visual
954 cognition, MIT press Cambridge, MA.
- 955
956 Yamins, D. L. and J. J. DiCarlo (2016). "Using goal-driven deep learning models to understand
957 sensory cortex." Nature neuroscience **19**(3): 356-365.
- 958
959 Yamins, D. L., et al. (2013). Hierarchical modular optimization of convolutional networks
960 achieves representations similar to macaque IT and human ventral stream. Advances in neural
961 information processing systems.

962
963 Yamins, D. L., et al. (2014). "Performance-optimized hierarchical models predict neural
964 responses in higher visual cortex." Proceedings of the National Academy of Sciences:
965 201403112.

966
967 Zeiler, M. D. and R. Fergus (2014). Visualizing and understanding convolutional networks.
968 Computer Vision—ECCV 2014, Springer: 818-833.

969

970

971

972 **TABLES**

973 **Table 1**

Behavioral Metric	Hit Rate	False Alarm Rate
One-versus all object-level performance (B.O1) (24 x 1) $O_1(i) = Z(HR(i)) - Z(FAR(i)),$ $i = 1, 2, \dots, 24$	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) $O_2(i, j) = Z(HR(i, j)) - Z(FAR(i, j)),$ $i = 1, 2, \dots, 24$ $j = 1, 2, \dots, 24$	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) $I_1(ii) = Z(HR(ii)) - Z(FAR(ii)),$ $ii = 1, 2, \dots, 240$ $j = 1, 2, \dots, 24$	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) $I_2(ii, j) = Z(HR(ii, j)) - Z(FAR(ii, j)),$ $ii = 1, 2, \dots, 240$ $j = 1, 2, \dots, 24$	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 .

974

975 **Table 1: Definition of behavioral performance metrics.** The first column provides the name,
 976 abbreviation, dimensions, and equations for each of the raw performance metrics. The next two
 977 columns provide the definitions for computing the hit rate (HR) and false alarm rate (FAR)
 978 respectively.

979

980 **FIGURE LEGENDS**

981

982 **Figure 1. Images and behavioral task. (A)** Two (out of 100) example images for each of the 24
983 basic-level objects. To enforce true invariant object recognition behavior, we generated
984 naturalistic synthetic images, each with one foreground object, by rendering a 3D model of each
985 object with randomly chosen viewing parameters and placing that foreground object view onto a
986 randomly chosen, natural image background. **(B)** Time course of example behavioral trial (zebra
987 versus dog) for human psychophysics. Each trial initiated with a central fixation point for 500
988 ms, followed by 100 ms presentation of a square test image (spanning 6-8° of visual angle).
989 After extinction of the test image, two choice images were shown to the left and right. Human
990 participants were allowed to freely view the response images for up to 1000 ms and responded
991 by clicking on one of the choice images; no feedback was given. To neutralize top-down feature
992 attention, all 276 binary object discrimination tasks were randomly interleaved on a trial-by-trial
993 basis. The monkey task paradigm was nearly identical to the human paradigm, with the
994 exception that trials were initiated by touching a fixation circle horizontally centered on the
995 bottom third of the screen, and successful trials were rewarded with juice while incorrect choices
996 resulted in timeouts of 1–2.5s. **(C)** Large-scale and high-throughput psychophysics in humans
997 (top left), monkeys (top right), and models (bottom). Human behavior was measured using the
998 online Amazon MTurk platform, which enabled the rapid collection ~1 million behavioral trials
999 from 1472 human subjects. Monkey behavior was measured using a novel custom home-cage
1000 behavioral system (MonkeyTurk), which leveraged a web-based behavioral task running on a
1001 tablet to test many monkey subjects simultaneously in their home environment. Deep
1002 convolutional neural network models were tested on the same images and tasks as those
1003 presented to humans and monkeys by extracting features from the penultimate layer of each
1004 visual system model and training back-end multi-class logistic regression classifiers. All
1005 behavioral predictions of each visual system model were for images that were not seen in any
1006 phase of model training.

1007 **Figure 2. Object-level comparison to human behavior. (A)** One-versus-all object-level
1008 performance (B.O1) metric for the pooled human (n=1472 human subjects), pooled monkey
1009 (n=5 monkey subjects), and several DCNN_{IC} models. Each B.O1 pattern is shown as a 24-
1010 dimensional vector using a color scale; each colored bin corresponds to the system's

1011 discriminability of one object against all others that were tested. The color scales span each
1012 pattern's full performance range, and warm colors indicate lower discriminability. **(B)** Direct
1013 comparison of the B.O1 metric vector computed from the behavioral output of a pixel visual
1014 system model (top panel) and a DCNN_{IC} visual system model (Inception-v3, bottom panel)
1015 against that of the human B.O1 metric vector. **(C)** Consistency to the human pool, with respect to
1016 the B.O1 metric, for each of the tested model visual systems. The black and gray dots correspond
1017 to a held-out pool of four human subjects and a pool of five macaque monkey subjects
1018 respectively. The shaded area corresponds to the "primate zone," a range of consistencies
1019 delimited by the estimated consistency of a pool of infinitely many monkeys (see Figure 4A).
1020 **(D)** One-versus-other object-level performance (B.O2) metric for pooled human, pooled
1021 monkey, and several DCNN_{IC} models. Each B.O2 pattern is shown as a 24x24 symmetric
1022 matrices using a color scale, where each bin (i,j) corresponds to the system's discriminability of
1023 objects i and j . Color scales similar to (A). **(E)** Consistency to the human pool, with respect to
1024 the B.O2 metric, for each of the tested model visual systems. Format is identical to (C).

1025 **Figure 3. Image-level comparison to human behavior.** **(A)** Schematic for computing B.I1n
1026 metric. First, the one-versus-all image-level metric (B.I1) is shown as a 240-dimensional vector
1027 (24 objects, 10 images/object) using a color scale, where each colored bin corresponds to the
1028 system's discriminability of one image against all distractor objects. From this pattern, the
1029 normalized one-versus-all image-level metric (B.I1n) is estimated by subtracting the mean
1030 performance value over all images of the same object. This normalization procedure isolates
1031 behavioral variance that is specifically image-driven but not simply predicted by the object. **(B)**
1032 Normalized one-versus-all object-level performance (B.I1n) metric for the pooled human, pooled
1033 monkey, and several DCNN_{IC} models. Each B.I1n pattern is shown as a 240-dimensional vector
1034 using a color scale, formatted as in (A). Color scales similar to Figure 2A. **(C)** Consistency to the
1035 human pool, with respect to the B.I1n metric, for each of the tested model visual systems. Format
1036 is identical to Figure 2C. **(D)** Normalized one-versus-other image-level performance (B.I2n)
1037 metric for pooled human, pooled monkey, and several DCNN_{IC} models. Each B.I2n pattern is
1038 shown as a 240x24 matrix using a color scale, where each bin (i,j) corresponds to the system's
1039 discriminability of image i against distractor object j . Color scales similar to Figure 2A. **(E)**
1040 Consistency to the human pool, with respect to the B.I2n metric, for each of the tested model
1041 visual systems. Format is identical to Figure 2C.

1042 **Figure 4. Effect of subject pool size and DCNN model modifications on consistency with**
1043 **human behavior. (A)** Accounting for natural subject-to-subject variability. For each of the four
1044 behavioral metrics, the human consistency distributions of monkey (blue markers) and model
1045 (black markers) pools are shown as a function of the number of subjects in the pool. The human
1046 consistency increases with growing number of subjects for all visual systems across all
1047 behavioral metrics. The dashed lines correspond to fitted exponential functions, and the
1048 parameter estimate (mean \pm SE) of the asymptotic value, corresponding to the estimated human
1049 consistency of a pool of infinitely many subjects, is shown at the right most point on each
1050 abscissa. **(B)** Model modifications that aim to rescue the DCNN_{IC} models. We tested several
1051 simple modifications (see Methods) to the DCNN_{IC} visual system model that scored the highest
1052 in our benchmarks (Inception-v3). Each panel shows the resulting human consistency per
1053 modified model (mean \pm SD over different model instances, varying in random filter
1054 initializations) for each of the four behavioral metrics.

1055
1056 **Figure 5. Analysis of unexplained human behavioral variance. (A)** Residual similarity
1057 between all pairs of human visual system models. The color of bin (i,j) indicates the proportion
1058 of explainable variance that is shared between the residual image-level behavioral patterns of
1059 visual systems i and j . For ease of interpretation, we ordered visual system models based on their
1060 architecture and optimization procedure and partitioned this matrix into four distinct regions. **(B)**
1061 Summary of residual similarity. For each of the four regions in Figure 5A, the similarity to the
1062 residuals of Inception-v3 (region 2 in (A)) is shown (mean \pm SD, within each region) for all
1063 images (black dots), and for images that humans found to be particularly difficult (blue and red
1064 dots, selected based on held-out human data).

1065
1066 **Figure 6. Dependence of primate and DCNN model behavior on object viewpoint and pixel**
1067 **attributes. (A)** Example images with increasing attribute value, for each of the six pre-defined
1068 image attributes. **(B)** Dependence of performance as a function of six image attributes, for
1069 humans, monkeys and a DCNN_{IC} model (Inception-v3). **(C)** Proportion of explainable variance
1070 of the residual image-level behavioral pattern of monkeys (black), an Inception-v3 model (dark
1071 blue), and all DCNN_{IC} models (light blue) that is accounted for by each of the pre-defined image

1072 attributes. Error-bars correspond to SD over trial re-sampling for monkeys, over different model
1073 “subjects” for Inception-v3, and over different DCNN_{IC} models for “Models (all)”.

1074

Figure 1

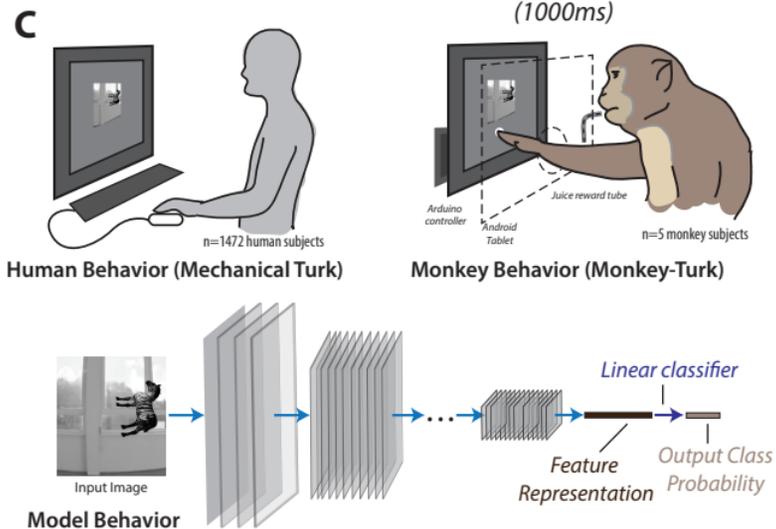
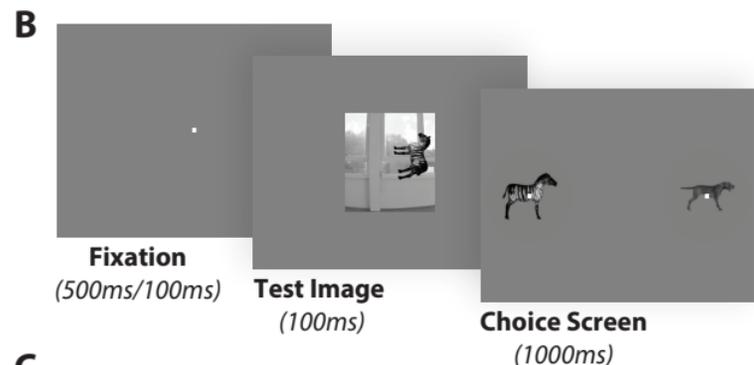
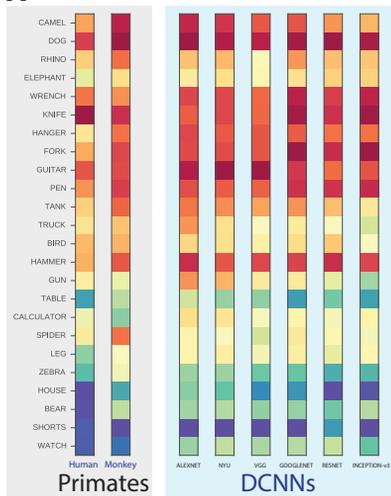
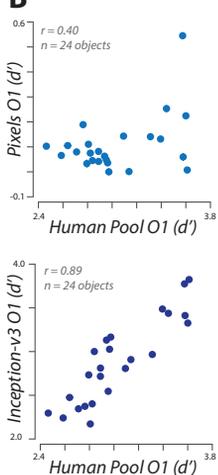


Figure 2

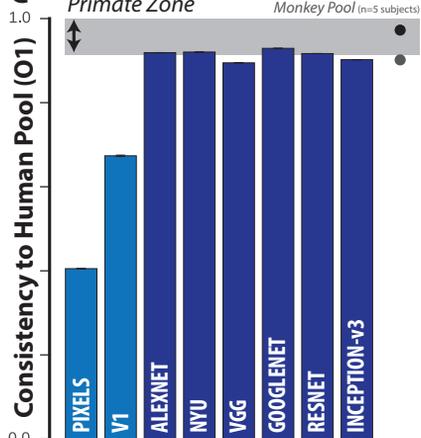
A B.O1 (one-vs-all object-level performance)



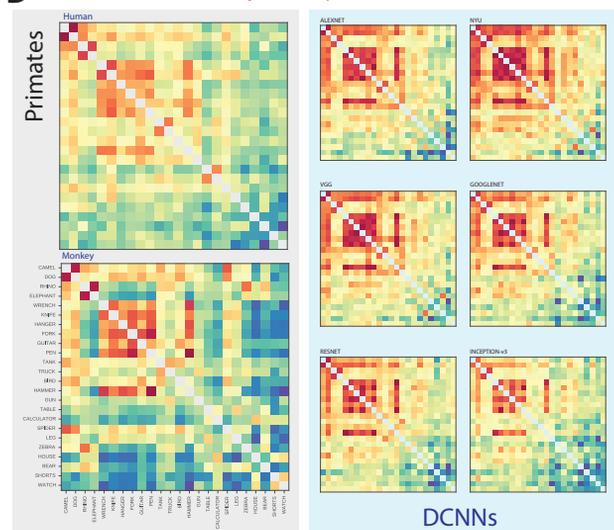
B



C



D B.O2 (one-vs-other object-level performance)



E

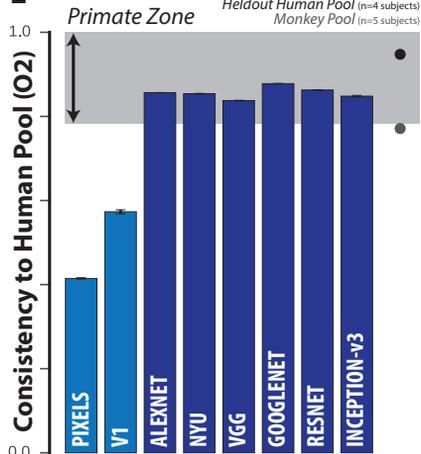
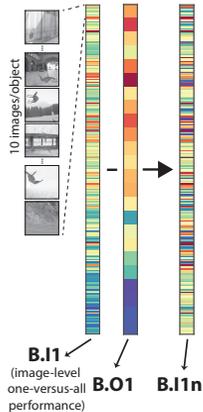
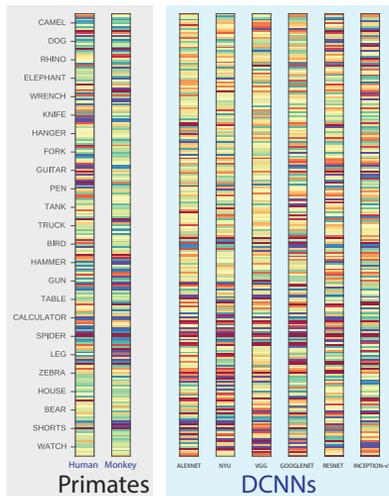


Figure 3

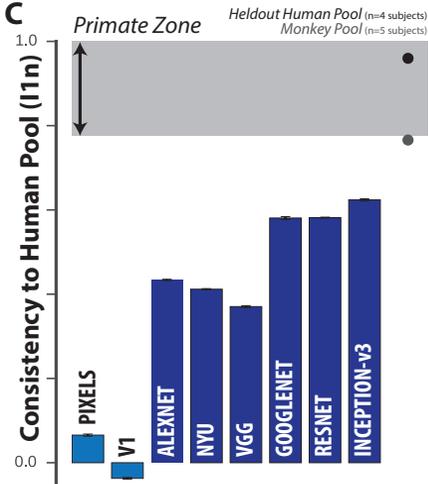
A B.11n (normalized one-vs-all image-level performance)



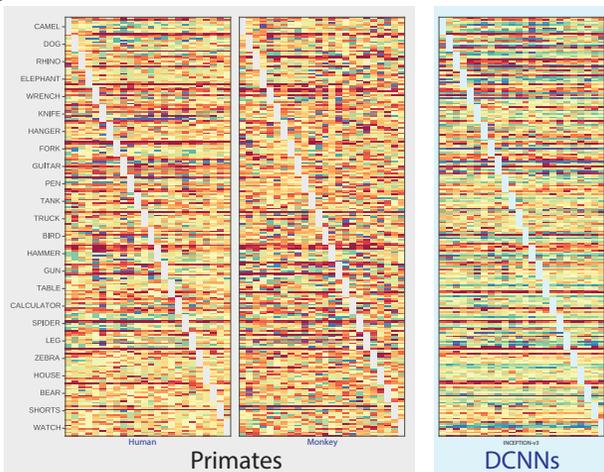
B



C



D B.12n (normalized one-vs-other image-level performance)



E

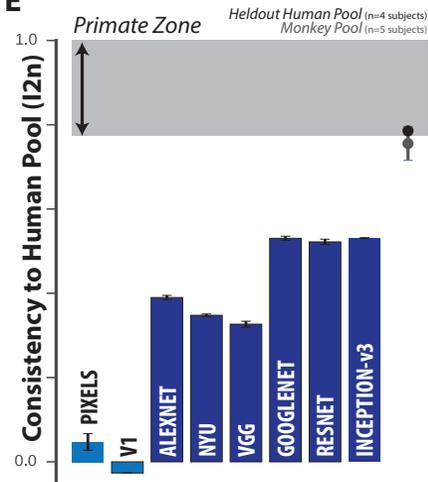


Figure 4

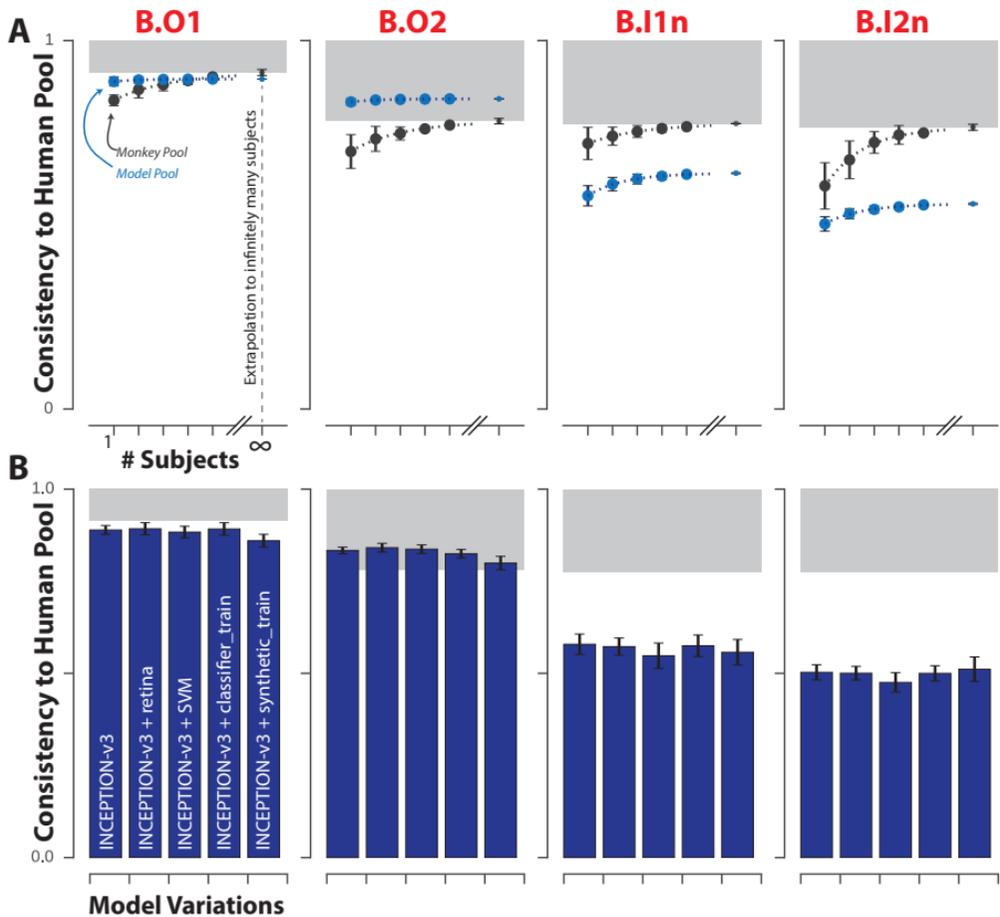


Figure 5

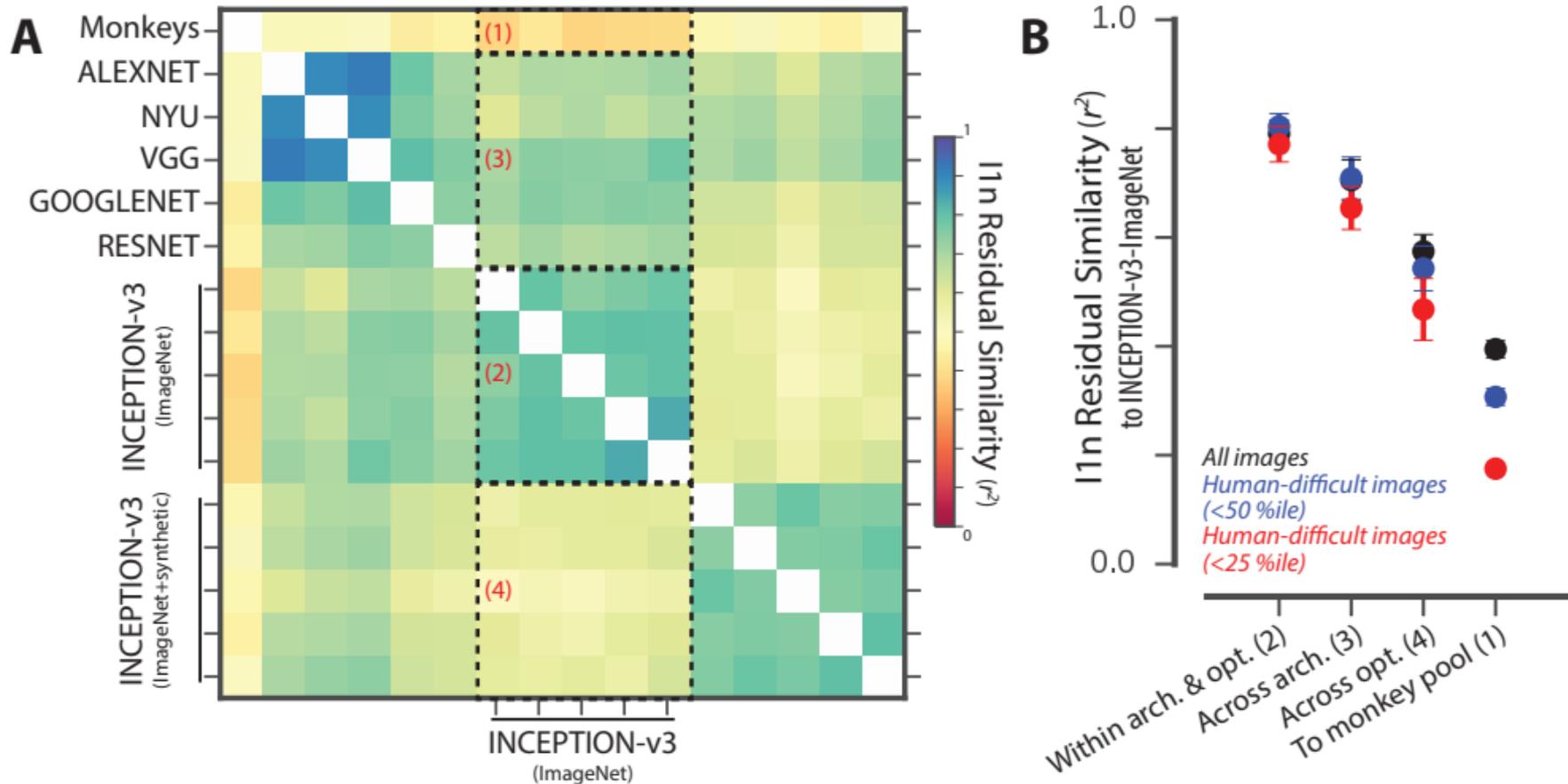
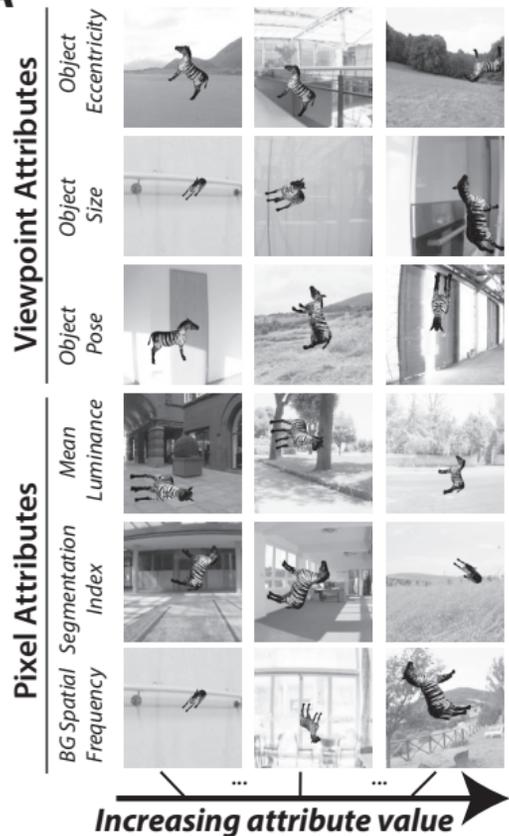
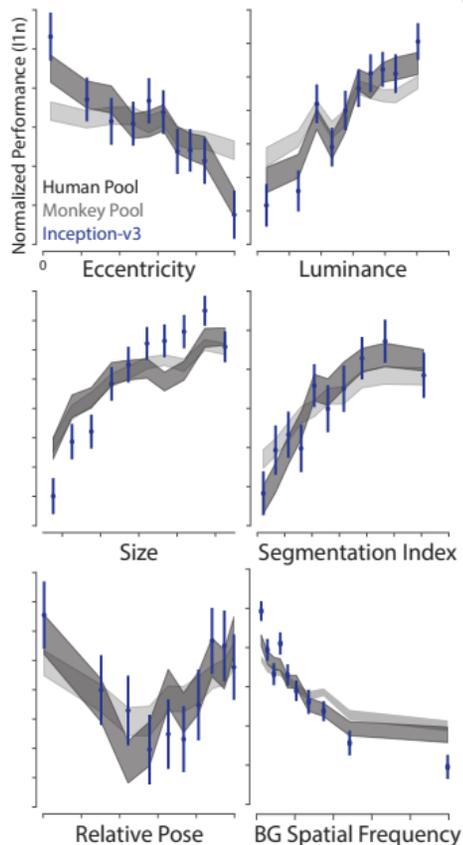


Figure 6

A



B



C

