Conserved core visual object recognition across simian primates: Marmoset image-by-image behavior mirrors that of humans and macaques

Alexander J.E. Kell[1,2,3,4], Sophie L. Bokor[1,2,3], You-Nah Jeon[1,2], Tahereh Toosi[1,2], Elias B. Issa[1,2,4]

1. Zuckerman Mind Brain Behavior Institute, Columbia University
2. Department of Neuroscience, Columbia University
3. Co-authors
4. Correspondence: alex.kell@columbia.edu, elias.issa@columbia.edu

We report that the marmoset, a 300-gram simian primate with a flat cortex, performs a challenging high-level perceptual task in a strikingly human-like manner. Across the same set of 400 images, marmosets' image-by-image core object recognition behavior was highly similar to that of humans—and was nearly as human-like as was macaques' (r=0.73 vs. r=0.77). Separately, we found that marmosets' visual abilities far outstripped those of rodents—marmosets substantially outperformed rats and generalized in a far more robust manner across images. Thus, core aspects of visual perception are conserved across simian primates, and marmosets may be a powerful small model organism for high-level visual neuroscience.

Descended from a common ancestor, the simian primates consist of monkeys, apes, and humans. Among the smallest of the simians, the common marmoset (*Callithrix jacchus*) was introduced to neuroscience decades ago[1-3] and has recently been adopted more broadly[4-13]. In contrast to larger simian primate model organisms like rhesus macaques, marmosets may mitigate traditional compromises between an animal model's methodological tractability and its phylogenetic proximity to humans, offering features typically associated with rodents in a simian—e.g., small size (~300 g), flat cortex, and relatively high fecundity (see Supp. Fig. 1). But it is not obvious how human-like marmosets may be in terms of their brain and behavior. Marmosets are a New World monkey, the simian clade most distant from humans. Moreover, even within the New Worlds, marmosets and their closest relatives were originally considered primitive and squirrel-like, distinct from the "true monkeys"[14-16]—in part because of the same attributes that make them an attractive animal model, their diminutive size and lissencephalic cortex. Thus, simian or not, the breadth of marmosets' utility as a small animal model for twenty-first century neuroscience remains unclear, particularly for high-level perception and cognition.

Here, we assessed the advanced perceptual abilities of the marmoset in a quintessential domain of systems neuroscience: vision. We quantitatively compared the behavior of three simian primates—marmosets, humans, and macaques—during core visual object recognition[17], the invariant identification of objects on natural backgrounds in the face of changes in object scale, position, and pose (Fig. 1a). We selected this behavior to benchmark the marmoset because it is challenging[18] and thought to capture a crux problem in natural vision[17,19]. We measured marmoset (n=5; n trials=205,667; average of 948 trials per day; Fig. 1c, Supp. Fig. 2, Supp. Video 1) and human (n=7; n trials=146,496) performance on each of the same 400 images, and compared this image-by-image behavior with previously collected macaque data[20], again on identical images. On each trial, marmosets, humans, and macaques chose which of two basic-level objects (e.g., camel or wrench) was present in a briefly flashed image (Fig. 1b, Supp. Fig. 3). We found that humans performed this two-alternative forced-choice task well, but by no means perfectly, achieving 94% accuracy (Fig. 1d), which demonstrates that these images comprise a challenging object recognition task.

We first evaluated whether marmosets could even perform this task. We measured marmoset monkeys' behavior on the identical set of 400 images and found that they performed at 80% accuracy. Mean accuracy per marmoset was 88%, 86%, 78%, 76%, and 70% (Fig. 1d). By contrast, we found that a linear classifier trained on image pixels performed at chance, demonstrating that this task required some degree of non-trivial visual processing. We employed no subject inclusion or selection criteria—and thus these individual performances are representative of marmosets' capabilities, rather than a reflection of a few outlier high-performing animals. The high performance of each subject demonstrates that this small New World primate performs well at a demanding high-level perceptual task.

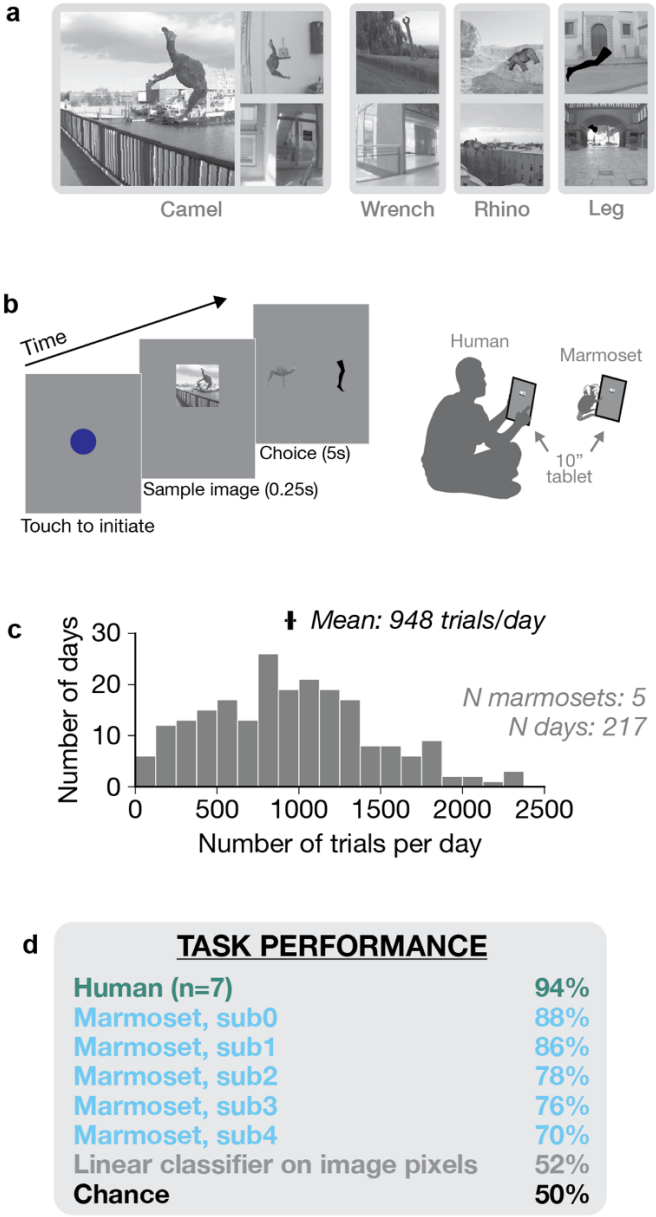


Figure 1. Invariant object recognition task and marmoset behavior. (a) Example images (of 400 total) from the four objects tested. **(b)** Task design. **(c)** Histogram of trials per day by marmosets. **(d)** Task performance for humans, each marmoset subject, and a baseline control model. While performance was measured on identical images, presentation time differed across marmosets and humans (250 and 50 msec, respectively).

Although marmosets performed the task well, it remained unclear whether they achieved this high performance in a manner similar to that of humans. Humans and marmosets could reach high performance through different means of visual processing, which would likely reveal themselves in different patterns of performance across the images. To assess this possibility, we compared marmoset and human visual object recognition behavior at a granular, image-by-image level. We computed the difficulty of each image and subtracted off the average difficulty of each object to yield a metric we refer to as the "i1n"[20] (Fig. 2a; see Methods for details). Unlike one-dimensional summaries of overall performance, this image-by-image 400-length vector is a rich, high-dimensional signature of visual perception that is robust to global, non-perceptual factors like attentional lapses or motor errors. Such

global factors lead to absolute shifts in performance but leave the relative image-by-image pattern intact. We compared marmoset and human i1ns, and found the two to be remarkably correlated ($r = 0.73$, $p = 8.91 \times 10^{-68}$; Fig. 2b, Supp. Fig. 5)—both species tended to find the same images easy or difficult. It was not obvious how to equate image size across humans and marmosets, as humans have ~2x the foveal acuity (~60 vs. 30 cycles per degree[21,22]), but marmosets exhibited human-like patterns when they performed the task at smaller or larger sizes spanning an octave (marmosets ~22° images, n trials = 98,378, $r = 0.65$, $p = 2.17 \times 10^{-49}$; Fig. 2c; marmosets ~11° images, n trials = 107,289: $r = 0.71$, $p = 1.35 \times 10^{-62}$; Fig. 2d). The high correlation between marmosets and humans is rendered even more impressive by the fact that different humans exhibited slightly different i1ns, even after accounting for measurement noise (mean cross-subject correlation, $r = 0.88$). These results suggest that marmosets—despite having both a brain and a body size 1/200th the mass of humans'—perform invariant visual object recognition in a strikingly human-like manner.

To further test the hypothesis that core recognition is common across simian primates, we compared the New World marmosets with an Old World monkey model of vision, the rhesus macaque, using macaque performance from a previous study on the identical 400 images[20]. Marmosets and macaques exhibited similar image-by-image behavioral signatures ($r = 0.80$, $p = 2.52 \times 10^{-90}$; Fig. 2e), and macaques were only slightly more similar to humans than were marmosets (macaque-human $r = 0.77$; marmoset-human $r = 0.73$, dependent t-test of correlation coefficients[23,24]: $t_{397} = 2.06$, $p = 0.040$). Some details of experimental design in the previous work differed from ours (e.g., whether objects were interleaved within or across sessions; see Methods for details), which raises the possibility that we may have underestimated the macaque-human similarity relative to the marmoset-human similarity, despite the use of identical images across all three species. It seems unlikely that these differences were particularly consequential for the i1n, as human signatures collected across the two settings were highly correlated ($r = 0.90$, $p = 1.32 \times 10^{-145}$). Moreover, the macaque-human similarity reported in the previous work, when macaques and humans were in more similar task settings, was comparable to what we estimated here across task settings (previous work[20], macaque-human $r = 0.77$; here, $r = 0.77$). The i1n therefore appeared to be primarily determined by the perceptual challenge of core object recognition, which was common across experimental settings. Critically, achieving this degree of similarity to humans is non-trivial, and not simply a consequence of high task performance[25]—artificial visual systems which performed at least as well as marmosets nonetheless exhibited i1ns which were substantially less human-like than marmoset i1ns (Fig. 2f; Supp. Fig. 4; AlexNet-humans $r = 0.52$, dependent t-test of correlation coefficients $t_{397} = 6.37$, $p = 5.29 \times 10^{-10}$; see Methods for details; also see[20]). Taken together, these results demonstrate that marmosets, humans, and macaques exhibit strikingly similar high-level visual behavior, and that, at least through the lens of image-by-image performance, marmosets' core object recognition is nearly as human-like as is macaques'.

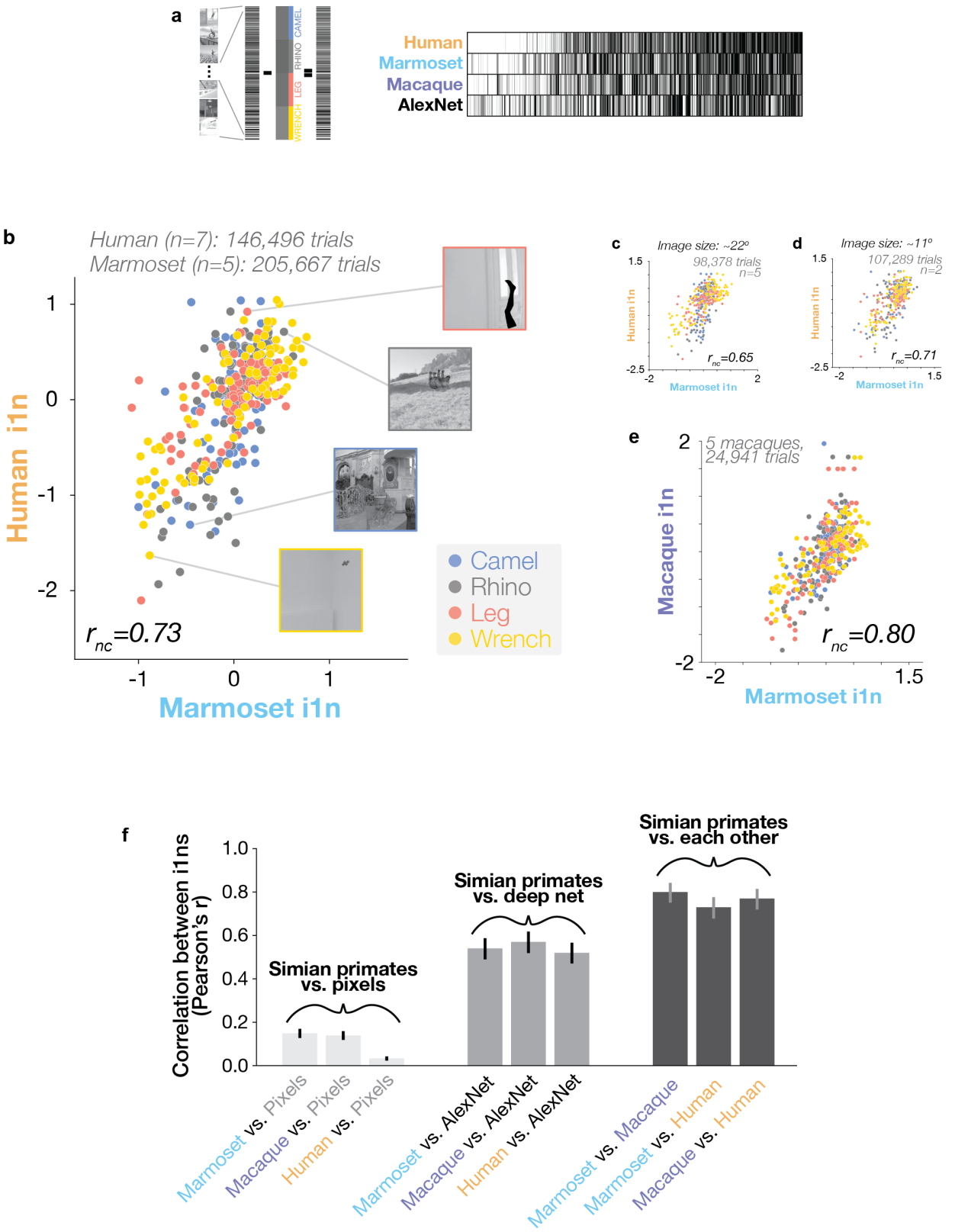


Fig. 2. Image-by-image behavioral comparisons of simian primates. (a) Image-by-image behavioral metric (i1n). Left: Normalization subtracts off object-level means, leaving the fine-grained, image-by-image pattern. Right: i1n signatures (barcodes) for marmosets, humans, macaques, and an artificial deep neural network (AlexNet), sorted by left-out human data. **(b)** Scatter comparing marmoset and human behavior for each of 400 images. Color indicates object; r_{nc} denotes noise-corrected correlation. Inset: Images for example points. **(c,d)** Scatter of marmoset behavior at, respectively, ~22° and ~11° degree image

size versus human behavior. **(e)** Scatter comparing marmoset and macaque behavior. **(f)** Pairwise correlation between simian primates' i1n performance signatures and those of pixel classifiers, a deep neural network, and each other. Error bars indicate 95% confidence intervals.

Finally, to zoom out and contextualize the marmoset's visual perceptual abilities, we compared marmoset visual behavior with that of another small animal model, the rat, as rodents have become an increasingly popular model for visual neuroscience[26,27] and prior work has quantified rat visual behavior in object recognition tasks[28]. We replicated one such task in marmosets, directly and quantitatively evaluating marmosets and rats on the identical stimuli and task, where animals were trained on a two-alternative forced-choice task to recognize two synthetic objects in isolation on black backgrounds and then tested on novel images of those same isolated objects under previously unencountered conjunctions of rotations and scales[28] (Fig. 3a). We found that marmosets performed substantially better than rats (mean accuracy marmosets: 93%, rats: 71%; two-tailed t-test: $t_{53} = 18.4$, $p = 1.14 \times 10^{-24}$; Fig. 3b and 3c). Moreover, marmosets and rodents exhibited qualitatively different patterns of generalization: rodents performed substantially worse across changes in rotation and scale, whereas marmosets were largely unaffected (Fig. 3d and 3e; ANOVA: species-by-scale interaction $F_{5,78} = 39.9$, $p = 3.81 \times 10^{-20}$, species-by-rotation interaction $F_{8,78} = 7.62$, $p = 1.78 \times 10^{-7}$). These differences were observed even though image size was scaled up to 40 degrees for rodents, in order to accommodate their coarse visual acuity (visual acuity of human, marmoset, and rat, respectively: 60, 30, and 1 cycles per degree[21,22,29]).

We then computationally benchmarked this rodent task, evaluating how well linear classifiers performed on the image pixels, in order to evaluate how well simple read-out mechanisms perform on the unprocessed visual input. For the core invariant recognition task originally used in macaques and humans, we had found that pixel-based linear classifiers performed at chance levels (near 50%, Fig. 1d). By contrast, for the rodent task we found that a linear classifier fit to the pixels of the animal's training images generalized nearly perfectly to the held-out images (97.5% accuracy). This result demonstrates that high performance on the rodent task does not require sophisticated representations for visual form processing. Therefore, despite previous interpretations[28], this paradigm may not be a strong behavioral test of invariant object recognition (also see[30]). Even on this relatively simple visual task, marmosets still out-performed rodents and generalized in a far more robust manner. Marmosets may therefore be a comparatively appealing small animal model of visual perception.

In summary, we found that marmosets exhibited human-like core visual object recognition behavior—indeed, on the identical set of images, marmosets were nearly as human-like as were macaques, the gold standard of visual systems neuroscience. Moreover, we found that marmosets' visual capabilities far outstripped those of rats, a common non-simian small animal model. Thus, aspects of core high-level perception appear to be shared across simian primates, and marmosets may offer a powerful platform for visual systems neuroscience that combines methodological advantages traditionally associated with

rodents (see Supp. Fig. 1) with high-level perceptual capabilities conventionally considered to be the exclusive purview of larger primates.

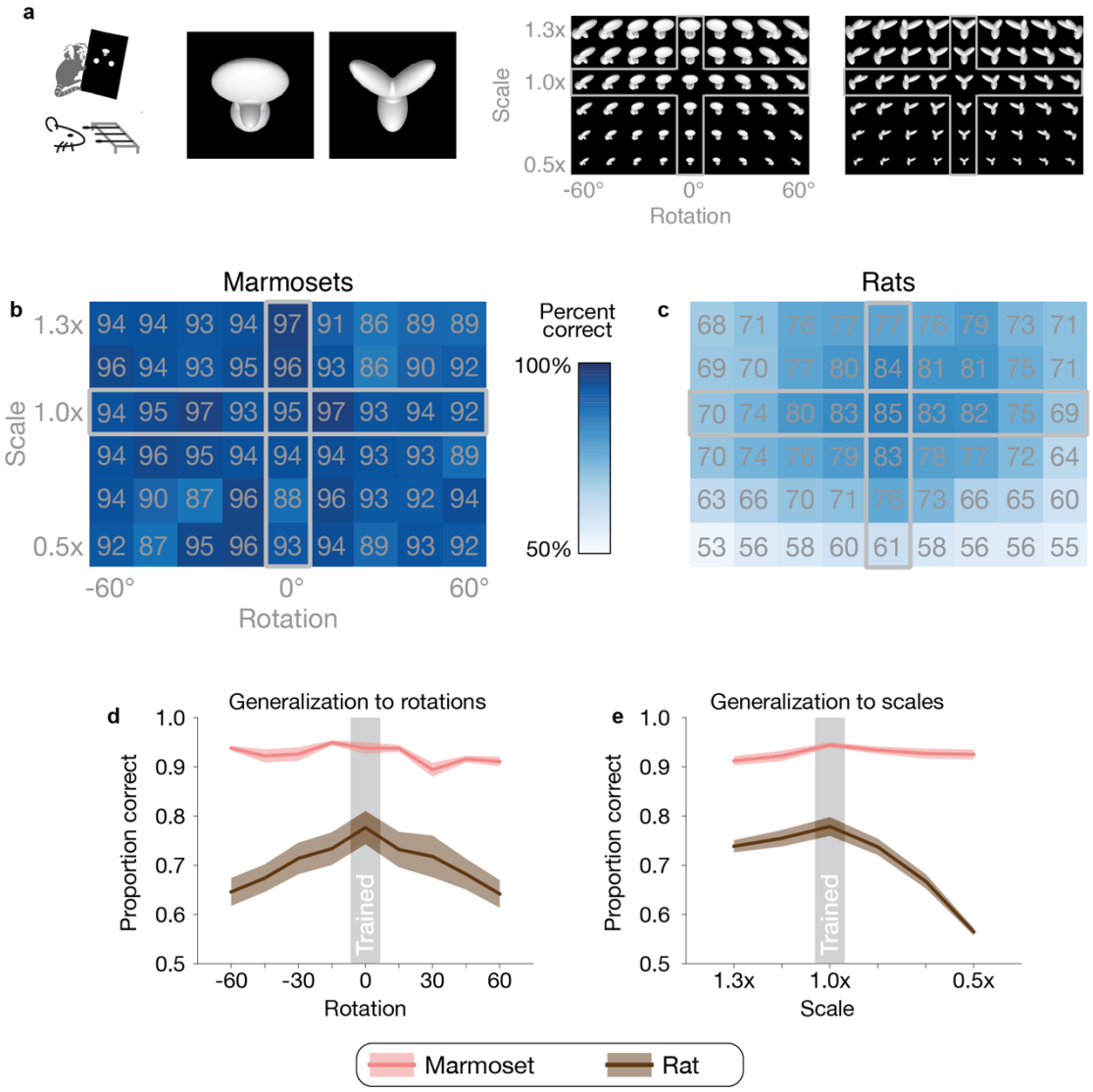


Fig. 3. Comparison to rodents. (a) The two objects used in a prior rodent study[25] (left) and all images generated by varying object rotation and scale (right). To test generalization, rats and marmosets were trained on a subset of images (outlined in gray) and then evaluated on all. **(b)** Marmoset accuracy at each rotation and scale. Overlaid: Percent correct for each image. Gray outline indicates images on which marmosets were trained. **(c)** Rat accuracy at each rotation and scale, reproduced from Zoccolan et al., 2009. Plotting conventions, including color scale, same as (b). **(d)** Generalization to novel images at each rotation. Both species were trained at 0°. Error bars are SEM over scales. **(e)** Generalization to novel images at each scale. Both species were trained at 1.0x. Error bars are SEM over rotations.

In this work, we not only showed that marmosets can perform a task as a proof-of-concept of their utility, but we also directly compared different species with identical stimuli and tasks, ultimately comparing to the high standard of human behavior. This kind of behavioral benchmarking allows characterizing the capacities of different species on equal footing. It enables identifying methodologically tractable animals that preserve a clearly defined behavioral phenomenon of interest—and therefore allows identifying as simple of animal model as possible, but no simpler. Behavioral benchmarking can be seen as part of a top-down, behavior-first approach to neuroscience[31], wherein the virtues of reductionism[32] are

complemented with real-world behavior, neuroethology, and, ultimately, biological intelligence. In this spirit, we hope that emerging small animal prosimian and non-primate models of high-level vision (e.g., mouse lemurs and tree shrews, respectively) will be quantitatively benchmarked. To foster this benchmarking, we are releasing all images, the marmoset and human behavioral data we collected, and as the macaque and rat data from previous work (github.com/issalab/kell-et-al-marmoset-benchmarking).

The similarity of high-level vision between marmosets, a 300-gram New World monkey, and macaques and humans may at first be surprising, but may make sense in the light of evolution. Recent large-scale genomic analyses show that macaques' and marmosets' last common ancestors with humans lived 32 and 43 million years ago, respectively (95% confidence ranges: 26-38 and 39-48 MYA)[33]. By contrast, the closest evolutionary relatives to humans outside of the simians are the prosimians (e.g., tarsiers, galagos, lemurs), and they most recently shared an ancestor with humans over 80 million years ago[33]. Key visual system features are thought to have evolved after this split and thus are specific to the simians, such as an all-cone, high-acuity fovea[21,34]. Moreover, a great elaboration of cortical areas appears to have occurred after this divergence as well—simians exhibit over one hundred cortical areas, whereas prosimians exhibit far fewer[35], and rodents fewer yet[36] (Supp. Table 1). Originally, marmosets' small size and flat cortex were interpreted as primitive—traits inherited without change from an ancestor—and thus in part a reflection of the phylogenetic gulf between marmosets and larger simians such as humans and macaques[37]. But their diminutive stature and flat cortex are now both understood to be derived, not primitive—marmosets' evolutionary ancestors have consistently shrunk over the past twenty million years[33,38], and their brains appear to have become increasingly lissencephalic over that time[39]. Moreover, rather than being an adaptation, the extent of gyrification across different species may simply be a consequence of the same geometrical principles that govern the crumpling of a sheet of paper[40]. All told, neither their small size nor flat brain are primitive, and marmosets may thus be thought of—in a literal sense—as a scaled-down simian primate.

As scaled-down simians, might marmosets exhibit other human-like core behaviors? Promising candidates include other aspects of high-level visual or auditory perception[41], as well as domains of motor behavior[5,7,12,42], as the corresponding regions of cortex have expanded the least between marmosets and humans[43]. Meanwhile, prefrontal, posterior parietal, and temporoparietal cortex, expanded the most[43], and thus in behaviors supported by these cortices, macaques may be more human-like[44]. However, even in these domains there may be exceptions potentially because of convergent evolution, such that marmosets exhibit human-like prosocial behaviors largely not seen in macaques[45]. Future work could behaviorally benchmark marmosets and other organisms in additional domains of perception and beyond, and in doing so, delineate homologies and convergences across taxa, while laying the

groundwork for understanding the neural mechanisms supporting real-world behaviors and the natural intelligence that underlies them.

References

1 Rosa, M. G. & Tweedale, R. Brain maps, great and small: lessons from comparative studies of primate visual cortical organization. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 665-691 (2005).

2 Lu, T., Liang, L. & Wang, X. Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neuroscience* **4**, 1131-1138 (2001).

3 Dias, R., Robbins, T. W. & Roberts, A. C. Dissociation in prefrontal cortex of affective and attentional shifts. *Nature* **360**, 69-72 (1996).

4 Mitchell, J. F., Reynolds, J. H. & Miller, C. T. Active vision in marmosets: a model system for visual neuroscience. *Journal of Neuroscience* **34**, 1183-1194 (2014).

5 Sedaghat-Nejad, E. *et al.* Behavioral training of marmosets and electrophysiological recording from the cerebellum. *Journal of Neurophysiology* **122**, 1502-1517 (2019).

6 Mitchell, J. F. & Leopold, D. A. The marmoset monkey as a model for visual neuroscience. *Neuroscience Research* **93**, 20-46 (2015).

7 Walker, J., MacLean, J. & Hatsopoulos, N. G. The marmoset as a model system for studying voluntary motor control. Developmental neurobiology. *Developmental Neurobiology* **77**, 273-285 (2017).

8 Gultekin, Y. B. & Hage, S. R. Limiting parental feedback disrupts vocal development in marmoset monkeys. *Nature Communications* **8**, 1-9 (2017).

9 Nurminen, L., Merlin, S., Bijanzadeh, M., Federer, F. & Angelucci, A. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nature Communications* **9**, 1-13 (2018).

10 Liao, D. A., Zhang, Y. S., Cai, L. X. & Ghazanfar, A. A. Internal states and extrinsic factors both determine monkey vocal production. *Proc Natl Acad Sci U S A* **115**, 3978-3983 (2018).

11 Oikonomidis, L. *et al.* A dimensional approach to modeling symptoms of neuropsychiatric disorders in the marmoset monkey. *Developmental Neurobiology* **77**, 328-353 (2017).

12 Ebina, T. *et al.* Two-photon imaging of neuronal activity in motor cortex of marmosets during upper-limb movement tasks. *Nature Communications* **9**, 1-16 (2018).

13 Sasaki, E. *et al.* Generation of transgenic non-human primates with germline transmission. *Nature* **459**, 523-527 (2009).

14 Martin, R. D. Goeldi and the dwarfs: the evolutionary biology of the small New World monkeys. *Journal of Human Evolution* **22**, 367-393 (1992).

15 Hershkovitz, P. *Living new world monkeys (Platyrrhini)*. (University of Chicago Press, 1977).

16 Sussman, R. W. & Kinzey, W. G. The ecological role of the Callitrichidae: a review. *American Journal of Physical Anthropology* **64**, 419-449 (1984).

17 DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415-434 (2012).

18 Krizhevsky, A., Sutskever, I. & Hinton, G. in *Advances in Neural Information Processing Systems.*

19 Pinto, N., Cox, D. D. & DiCarlo, J. J. Why is real-world visual object recognition hard? *PLoS Computational Biology* **4** (2008).

20 Rajalingham, R. *et al.* Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* **38**, 7255-7269 (2018).

21 Kirk, E. C. & Kay, R. F. in *Anthropoid Origins: New Visions* (ed C. F. & Kay Ross, R. F.) 539–602 (Springer US, 2004).

22 Nummela, S. U. *et al.* Psychophysical measurement of marmoset acuity and myopia. *Developmental Neurobiology* **77**, 300-313 (2017).

23 Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* **87**, 245 (1980).

24 Williams, E. J. The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)* **21**, 369-399 (1959).

25 Kell, A. J. E. & McDermott, J. H. Deep neural network models of sensory systems: windows onto the role of task constraints. *Current opinion in neurobiology* **55**, 121-132 (2019).

26 de Vries, S. E. *et al.* A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience* **23**, 138-151 (2019).

27 Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361-365 (2019).

28 Zoccolan, D., Oertelt, N., DiCarlo, J. J. & Cox, D. D. A rodent model for the study of invariant visual object recognition. *Proc Natl Acad Sci U S A* **106**, 8748-8753 (2009).

29 Prusky, G. T., West, P. W. & Douglas, R. M. Behavioral assessment of visual acuity in mice and rats. *Vision Research* **40**, 2201-2209 (2000).

30 Vinken, K. & Op de Beeck, H. Deep Neural Networks Point to Mid-level Complexity of Rodent Object Vision. *bioRxiv* (2020).

31 Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480-490 (2017).

32 Rust, N. C. & Movshon, J. A. In praise of artifice. *Nature Neuroscience* **8**, 1647 (2005).

33 Perelman, P. *et al.* A molecular phylogeny of living primates. *PLos Genetics* **7**, e1001342 (2011).

34 Williams, B. A., Kay, R. F. & Kirk, E. C. New perspectives on anthropoid origins. *Proc Natl Acad Sci U S A* **107**, 4797-4804 (2010).

35 Kaas, J. H. in *Progress in Brain Research* (ed M.A. Hofman) Ch. 3, 61-81 (Elsevier, 2019).

36 Van Essen, D. C. *et al.* Cerebral cortical folding, parcellation, and connectivity in humans, nonhuman primates, and mice. *Proc Natl Acad Sci U S A* **116**, 26173-26180 (2019).

37 Ford, S. M. Callitrichids as phyletic dwarfs, and the place of the Callitrichidae in Platyrrhini. *Primates* **21**, 31-43 (1980).

38 Montgomery, S. H. & Mundy, N. I. Parallel episodes of phyletic dwarfism in callitrichid and cheirogaleid primates. *Journal of Evolutionary Biology* **26**, 810-819 (2013).

39 Kelava, I. *et al.* Abundant occurrence of basal radial glia in the subventricular zone of embryonic neocortex of a lissencephalic primate, the common marmoset Callithrix jacchus. *Cerebral Cortex* **22**, 469-481 (2012).

40 Mota, B. & Herculano-Houzel, S. Cortical folding scales universally with surface area and thickness, not number of neurons. *Science* **349**, 74-77 (2015).

41 Song, X., Osmanski, M., Guo, Y. & Wang, X. Complex pitch perception mechanisms are shared by humans and a New World monkey. *Proc Natl Acad Sci U S A* **113** (2016).

42 Johnston, K. D., Barker, K., Schaeffer, L., Schaeffer, D. & Everling, S. Methods for chair restraint and training of the common marmoset on oculomotor tasks. *Journal of Neurophysiology* **119**, 1636-1646 (2018).

43 Chaplin, T. A., Yu, H.-H., Soares, J. G. M., Gattass, R. & Rosa, M. G. P. A Conserved Pattern of Differential Expansion of Cortical Areas in Simian Primates. *Journal of Neuroscience* **33**, 15120-15125 (2013).

44 Preuss, T. M. Critique of pure marmoset. *Brain, Behavior, and Evolution* **92**, 92-107 (2019).

45 Burkart, J. M. *et al.* The evolutionary origin of human hyper-cooperation. *Nature Communications* **5**, 1-9 (2014).

46 Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology* **8**, 382-439 (1976).

47 Brunton, B., Botvinick, M. & Brody, C. Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95-98, doi:doi:10.1126/science.1233912 (2013).

48 Poddar, R., Kawai, R. & Ölveczky, B. P. A Fully Automated High-Throughput Training System for Rodents. *PLoS ONE* **8**, e83171 (2013).

49 Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience* **22**, 974-983 (2019).

50 Spearman, C. The proof and measurement of the association between two things. *The American Journal of Psychology* **15**, 72-101 (1904).

51 DiCarlo, J. J. & Johnson, K. O. Velocity Invariance of Receptive Field Structure in Somatosensory Cortical Area 3b of the Alert Monkey. *Journal of Neuroscience* **19**, 401-419 (1999).

52 Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357-362 (2020).

53 Jones, E., Oliphant, T. & Peterson, P. *SciPy: Open source scientific tools for Python*, <http://www.scipy.org> (2001).

54 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** (2011).

55 Yamins, D. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619-8624, doi:10.1073/pnas.1403112111 (2014).

56 Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology* **10**, e1003915, doi:doi:10.1371/journal.pcbi.1003915 (2014).

57 A, P. *et al.* in *Advanced in neural information processing systems* 8026-8037 (2019).

58 Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv* **arXiv:1805.12177** (2018).

59 Wong, P. & Kaas, J. H. Architectonic subdivisions of neocortex in the tree shrew (Tupaia belangeri). *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology* **292**, 994-1027 (2009).

60 Saraf, M. P. *et al.* Architectonic features and relative locations of primary sensory and related areas of neocortex in mouse lemurs. *Journal of Comparative Neurology* **527**, 625-639 (2019).

61 Zilles, K., Palomero-Gallagher, N. & Amunts, K. Development of cortical folding during evolution and ontogeny. *Trends in Neurosciences* **36**, 275-284 (2013).

62 Ventura-Antunes, L., Mota, B. & Herculano-Houzel, S. Different scaling of white matter volume, cortical connectivity, and gyrification across rodent and primate brains. *Frontiers in Neuroanatomy* **7** (2013).

63 Veilleux, C. C. & Kirk, E. C. Visual acuity in mammals: effects of eye size and ecology. *Brain, Behavior, and Evolution* **83**, 43-53 (2014).

Author Contributions

A.J.E.K., S.L.B., and E.B.I designed the research. S.L.B., Y.J, T.T., and E.B.I. developed the experimental the apparatus. A.J.E.K., S.L.B., Y.J., TT., and E.B.I. collected the data. A.J.E.K. analyzed the data. A.J.E.K. and E.B.I. wrote the manuscript. All authors reviewed and edited the manuscript. E.B.I. acquired funding and supervised the research.

Acknowledgements

The authors thank Robert Desimone, James DiCarlo, and Guoping Feng for early project support; Hector Cho, Elizabeth Yoo, and Michael Li for technical support; James DiCarlo and Rishi Rajalingham for the macaque data; Davide Zoccolan for providing the original images used in the rat behavior study; and Hector Cho, Aniruddha Das, Nancy Kanwisher, Jack Lindsay, Rishi Rajalingham, and Erica Shook for comments on the manuscript. The work was funded by an NIH Postdoctoral NRSA fellowship to A.K. (F32 DC017628), an NIH R00 to E.I. (EY022671), and a Klingenstein-Simons Fellowship in Neuroscience to E.I.

Methods

Subjects. Five common marmosets (*Callithrix jacchus*) and seven humans participated in our experiments. The five authors were among the participants. The human data were collected in accordance with the Institutional Review Board of Columbia University Medical Center, and the marmoset data were collected in accordance with the NIH guidelines and approved by the Columbia University Institutional Animal Care and Use Committee (IACUC). We also used data from two previously published studies[20,28] from five macaques, 1,481 humans, and six rats. These data were collected in accordance with NIH guidelines, the Massachusetts Institute of Technology Committee on Animal Care, The Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (COUHES), and the Harvard Institutional Animal Care and Use Committee.

Two-way stimulus-response task. Marmoset, human, and macaque performance was measured on the identical set of images (same object, pose, position, scale, and background). See Supplemental Figure 3 for these 400 images. Humans, macaques, and marmosets initiated each trial by touching a dot at the center of the screen. A sample image was then flashed (for marmosets 250 msec and for humans 50 msec; we varied presentation duration to induce a sufficient number of errors to yield reliable human image-by-image performance scores; the macaque data that was collected in previous work presented images for 100 msec). After the image disappeared, two example token images were presented (i.e., of the object on a gray background at a canonical viewing angle), and the subject had to touch which of two images was in the preceding sample image. We used a two-alternative forced-choice stimulus-response paradigm—for a given task session, whether for marmoset or human, only a pair of objects was tested (e.g., one day would be camel-wrench, another would be wrench-leg). This was varied across sessions so that images of a given object were tested in all possible two-way task settings (e.g., camel vs. rhino, camel vs. wrench, and camel vs. leg). For a given behavioral session, each object was consistently on the same side of the decision screen (e.g., camel would always be on left and wrench always on right). One motivation for using this kind of two-way stimulus-response design is that it reduces the working memory demands on the participant. As soon as the participant recognizes what is in the image, they can plan their motor action. In part because of this simplicity, the two-way stimulus response paradigm is broadly used across the neurosciences, including in rodents[28], and it therefore allows behavioral benchmarking for comparative studies across an array of candidate model species. Moreover, as reported in the main text, we found that the pattern of image-by-image errors (i1n) was highly similar across humans who performed the two-way stimulus-response design and those that performed the match-to-sample design, in which pair-wise image discriminations were interleaved within a session rather than across sessions ($r = 0.90$; this fully-interleaved match-to-sample design was used in the previous work with the macaques). The similarity of i1ns across experimental designs suggests that when

comparing model systems on purely perceptual grounds, a simplified, two-alternative stimulus-response task can serve as a universal method for comparative studies across animals—even those where a challenging match-to-sample task design may be somewhat difficult to employ.

For the rodent task, the design was similar to that of the core recognition task, but sample stimuli were not flashed and were instead present on the screen until the subject made their choice. We used this design because it mimicked Zoccolan and colleagues' study. Instead of requiring the marmosets to use a nose-poke into two ports, we had the marmosets touch one of two white circles on-screen, one on the left and one on the right, to indicate their choice. We used white circles rather than token images to avoid allowing the marmosets to employ a pixel-matching strategy with the image on the screen.

Stimuli. For the core object recognition task in marmosets, we used a subset of images from a previous study in macaques (see Supplemental Figure 3 for all 400 images). We briefly describe the image generation process below, but see Rajalingham et al., 2018 for more details[20]. These stimuli were designed to examine basic-level[46], core object recognition. They were synthetically generated, naturalistic images of four objects (camel, wrench, rhino, and leg) displayed on a randomly chosen natural image background. The four objects were a randomly selected subset of the twenty-four used in the previous work. The spatial (x, y) position of the center of the object, the two angles parameterizing its pose (i.e., 3d rotation), and the viewing distance (i.e., scale) were randomly selected for each image. Stimuli were designed to have a high degree of variation, with disparate viewing parameters and randomized natural image backgrounds, in an effort to capture key aspects of the challenge of invariant object recognition, and to remove potential low level confounds that may enable a simpler visual system, either biological or artificial, to achieve high performance[19]. In addition to this main set of 400 test images (100 for each object), we employed four additional sets of images in training the marmosets to do the task (for training details, see "Marmoset training procedure" below). The first consisted of a single token image: the object rendered at the center of a gray background in a relatively canonical pose (e.g., side view of an upright camel). The second set consisted of 400 images of each object at random poses, positions, and scales, all on a uniform, gray background. The third set consisted of a different sample of images that were drawn from the same distribution of generative parameters as our 400 test images (i.e., variation in pose, position, and scale on randomly selected natural backgrounds). To assess the generality of our image-by-image performance scores, we collected marmoset behavior to the test images at two different sizes. Marmoset position and distance from the viewing screen was not strictly controlled but was nonetheless relatively stereotyped (see "Homecage testing boxes" below), and the two image sizes subtended ~11° and ~22° of visual angle. For humans, we collected a single image size. Humans were not constrained in how they held the tablets on which the images were displayed, and this image size subtended ~4-12 degrees of visual angle, depending on how the tablet was held.

For comparing marmosets to rodents, identical images were used as in a prior study in rats; see Fig. 3a for all images, and see Zoccolan et al., 2009 for more details[28]. In brief: these stimuli were of one of two synthetic, artificial objects, which were rendered at the center of the screen on a uniform black background. Objects were rendered at one of nine different azimuthal rotations (-60° to +60°, in steps of 15°) and one of six different scales, spanning ~1.4 octaves of size (from 0.5x to 1.3x the size of a template; for marmosets, 1.0x subtended ~11 degrees visual angle). Rotations and scales were fully crossed for each object to yield a total of 108 images (9 rotations x 6 scales x 2 objects).

Web-based, homecage behavioral training system. In part inspired by the high-throughput behavioral systems used in some rodent work[47,48], we developed a system where behavior would be collected in parallel from a large number of animals.

Web-based behavioral platform. We developed the MkTurk web-based platform (***mkturk.com***) to collect the data on tablets that could be deployed anywhere. We opted for a web-based system for a variety of reasons. First, MkTurk just needs a web browser to run, and as a result the setup and installation across tablets is relatively low cost, both in terms of the researcher's time as well as money, as consumer touchscreen tablets are often cheaper than more traditional behavioral rigs. Second, such a system made it relatively turnkey to evaluate humans and marmosets in as similar environment as possible—we simply distributed the same tablets to humans, who performed the touchscreen tasks just as marmosets did. Third, being based on the web naturally enables real-time streaming of the animals' performance to automatic analysis pipelines in the cloud, which allows seamless monitoring of the animals on any device that can access that web (e.g., a researcher's smartphone). As a result, monitoring animal performance and troubleshooting issues in real-time is more straightforward. Moreover, since task parameters are passed from the cloud to the tablets in real time, task parameters can be adjusted on the fly remotely, which was particularly helpful during training of many subjects in parallel.

For external hardware, we leveraged the open-source Arduino platform (Arduino Leonardo) coupled to a low-power piezoelectric diaphragm pump for fluid delivery (Takasago Fluidic Systems) and RFID reader for individual animal identification (ID Innovations), all powered by a single 5V USB battery pack. Thus, our system may not be as powerful as some traditional, fully-equipped experimental rigs, but our behavior box was highly optimized for SWaP-C (Size ~= 1 ft^3, Weight ~= 10 lbs, Power ~= 5W, and Cost < $1,000).

Marmoset homecage behavior collection. Marmosets were tested in an 8" x 9" x 11" (width x height x depth) modified nest box that was attached to their housing unit. They were granted access to this box

three hours a day. Marmosets could move freely in and out of the testing box for *ad libitum* access to food during the three hours, and marmosets could engage with cagemates in social behavior when not performing the task. Marmosets performed trials on touchscreen tablets (Google Pixel C) inserted into a vertical slot in the rear of the nest box. Once in the nest box, marmoset body and head position were not strictly controlled (e.g., via chairing or head-fixing), but marmosets were encouraged into a relatively stereotyped position in a few ways. First, the boxes were augmented with dividers on both the left and right sides to restrict degrees of lateral freedom while marmosets were performing the task. Second, access to the touch screen was only available through a small (3.5" by 1.5") armhole in the front plexiglass barrier. Third, a metal reward tube was embedded in the front plexiglass barrier and positioned at the center of the screen. Given that marmosets tended to perform well, they received regular rewards of sweetened condensed milk via this tube, and they tended to position themselves at this tube consistently, yielding a relatively stereotyped viewing position (e.g., see Supplemental Movie 1). By collecting behavior from animals in their housing unit in the colony, we were able to acquire data from all marmosets simultaneously. Indeed, in part because of this high-throughput behavioral system, we were able to measure performance for each of 400 images with high precision by collecting many trials per image (trial split-half correlation for marmoset's i1n, Pearson's r = 0.94).

Human behavior collection. Human data was collected on the same tablet hardware using the same MkTurk web app software. As with the marmosets, humans were allowed to perform trials in their home environments. Moreover, we also measured human image-by-image performance with high precision and reliability (trial split-half correlation for human i1n, Pearson's r = 0.91).

Marmoset training procedure. Marmosets were trained on the task through a series of stages. The first stage aimed to familiarize the marmosets with the motor mapping. In this case the sample image that was flashed was simply one of the token images, and thus was the same as the images used as buttons in the decision stage. This task therefore required no visual processing beyond pixel-matching, but was helpful to familiarize marmosets with the task paradigm. In the second stage, we introduced random position, pose and scale, but still presented the images on plain gray backgrounds, and this task served as a helpful bridge between the simple motor-mapping component and the full-scale core object recognition task that was the goal of our training. In the third and final training stage, marmosets were trained on the high-variation images with disparities in position, pose, and scale and with natural-image backgrounds. When marmosets saturated performance on this final stage, we switched them to test images, and we report the behavior from these separate test images not used in the training stages.

Behavioral metric: i1n. *Definition of i1n*. To compare human and marmoset core object recognition, we employed an image-by-image metric, the i1n, which has been shown in previous work to be highly

discerning between visual systems[20] and a useful tool for selecting behaviorally challenging stimuli for studying neural processing in ventral visual cortex[49]. The i1n measures the discriminability of each image, and is designed to be a metric of a system's sensitivity, as formalized by signal detection theory. For each image *j*, the difficulty V_j is defined as:

$$V_j = z(HR_j) - z(FAR_j),$$

where *V* is a 400-length vector. HR_j and FAR_j are, respectively, the hit rate and the false alarm rate—the proportion of trials that image *j* was correctly classified and the proportion of trials that any image was incorrectly classified as that object. *z( )* is the inverse cumulative density function of a Gaussian distribution, which allows evaluating the difference between the hit rate and false alarm in z-scored units. For images where a system was correct on every trial, the z-transformed hit rate is infinity, and we capped these values to be 4, as this value was just outside the ceiling of our empirically measured hit rates given the number of trials that were in each bin. No images for marmosets nor humans reached this threshold; nine of the four hundred macaque images reached this threshold. We then subtracted off the mean value for each object to yield the i1n. Normalizing these values but subtracting off object-level performance (the "n" in the "i1n"), makes this metric robust to mean shifts in performance for each object and helps remove differential performance across behavior sessions. In a stimulus-responses design, object correlates with session. By contrast, when trials of different pairwise discriminations are interleaved within a session like in a match-to-sample design, any day-by-day variability in performance is not correlated with object-level performance. In practice, whether or not this object-level normalization was included did not affect the main results—Supplementary Figure 5 shows that key results are essentially the same when using an "i1" instead of an i1n.

Contextualizing the i1n. While the calculation of the metric was identical in our work and in the work by Rajalingham and colleagues, differences in the inputs to the i1ns may lead to some mild differences between the nature of each of our i1ns. First, our i1n was computed with more images. Rajalingham and colleagues reported human-macaque correlations of image-level metrics on a subset of 240 of the 2400 images tested (10 of 100 images of each of the 24 objects), because human data was somewhat expensive to acquire and, image-level metrics require large amounts of data per image to be reliably estimated. In our work, because we concentrated data collection on four objects, we collected many trials on 100 images of each object, and thus computed i1n from 400 images total. The consequences of different numbers of images are probably not particularly substantial, though high correlation coefficients are somewhat less likely between 400-dimensional than with 240-dimensional vectors (e.g., the standard deviation of a null distribution of correlations coefficients between random Gaussian 240- and 400-dimensional vectors are, respectively, 0.065 and 0.050). A second, and potentially more consequential, difference between our i1n and the i1n used by Rajalingham and colleagues is that they measured the performance of each image against twenty-three different distractor objects, whereas we only measured

ours against three. By averaging over far more distractor objects, their i1n likely minimizes the effect of the choice of distractor much more than our i1n does. Our i1n therefore likely measures something between Rajalingham and colleague's i1n and their i2n, which averages over no distractors. Nonetheless, these differences between our i1n and the i1n used in previous work may not lead to substantial differences. As reported in the main text, we find that our 400-dimensional macaque i1n (averaged over three distractors) and their 240-dimensional macaque i1n (averaged over 23 distractors) are both equally similar to the corresponding human i1ns collected in comparable situations (r_{nc}(macaque,human) = 0.77 in both cases).

Comparing i1ns: correcting correlation coefficients for test-retest reliability. To assess the similarity of different visual systems, we measured the consistency of the image-by-image performance of each by correlating i1ns (Pearson's correlation coefficient). We subjected the "raw" correlation coefficients to a correction that accounts for the test-retest reliability of the data, as different systems' i1ns will have different test-retest reliability due to, for instance, different amounts of data and/or different rates of errors. Reliability correction addresses potentially undesirable properties of comparing raw, uncorrected coefficients across pairs of systems. For instance, the consistency of a system with itself should be at ceiling (i.e., a correlation coefficient of 1), but an uncorrected correlation coefficient will have a value less than one; it will be determined by the test-retest reliability. Because of this, the ceiling would in general be different across pairs of systems, and left uncorrected, this could lead to inaccurate inferences—e.g., one could naively conclude that system 1 and system 2 are more similar than system 1 and system 3, just because the researchers measured more reliable i1ns in system 2 than in system 3. To address these concerns, we measured and corrected for the test-retest reliability of the i1n for each system, by applying the correction for attenuation[50,51], which estimates the noiseless correlation coefficient between the two—i.e., the correlation coefficient that would be observed as the number of trials goes to infinity. In doing so, we ensured that all comparisons between pairs of systems were on the same scale—i.e., the ceiling for each was indeed a correlation coefficient of 1—and thus were free to compare i1ns across these different systems.

We measured the noise-corrected correlation for a pair of systems' i1ns by randomly partitioning trials for each image into two halves, computing i1ns for each half for each system, taking the mean of the correlation coefficients of i1ns across systems across the two split halves, and dividing it by the geometric mean of the reliability across systems (this denominator being the correction for attenuation[50]):

$$R_{nc} = \frac{\frac{1}{2}\Big(r(V_{a0}, V_{b1}) + r(V_{a1}, V_{b0})\Big)}{\sqrt{r(V_{a0}, V_{a1})\ r(V_{b0}, V_{b1})}},$$

where R_{nc} denotes the noise-corrected correlation coefficient, $r(\ )$ is a function that returns the Pearson correlation coefficient between its two arguments, V_{a0} and V_{a1} denote splits of trials for system a, and V_{b0} and V_{b1} denote splits of trials for system b. For each comparison, to mitigate variation due to how the data was randomly partitioned, we took the mean of this noise-corrected correlation coefficient across 1000 random partitions.

To test whether the correlation coefficients between two pairs of systems were different (e.g., marmoset-human correlation versus macaque-human correlation), we employed a dependent t-test of correlation coefficients[23,24], which takes into account the dependency of the two correlation coefficients as they share a common variable (in this case humans). Accounting for this dependence increases the statistical power of the test.

Data analysis was conducted in Python and made use of the numpy[52], scipy[53], and sklearn[54] libraries.

Classifiers: Deep neural networks and pixels. To contextualize the behavior of simian primates, we also evaluated the overall performance and image-by-image performance for a variety of artificial systems. We trained linear classifiers for our task on top of representations from deep neural networks, which have been shown to be powerful models of visual behavior and neurophysiology[20,55,56]. We evaluated standard ImageNet-trained deep networks, downloading pretrained models from the torchvision package of PyTorch[57]. As a low-level control, we also compared performance of linear classifiers trained on image pixels (256 x 256 pixel images), which in part assesses the extent to which features like luminance and contrast covary with labels and thus can be used to perform the task.

Classifiers for task performance. We evaluated linear classifiers on the penultimate layer of deep networks, training linear support vector machines (SVMs) on the same images and same binary tasks that the primates performed. We used a hinge loss and L2 regularization. The features were nonnegative, because they were the activations of rectified linear units in the network, and so we did not center the data as that would alter the consequences of L2 regularization We instead experimented with allowing the model to learn an intercept term or not, and observed similar results in both cases; we report results when learning an intercept. To select the strength of regularization, we searched over a range of 21 logarithmically spaced hyperparameters and selected hyperparameter values via 80-20 splits within the training set. To mitigate variation of results due to which images were selected for training or testing, we trained 1000 classifiers on random train-test partitions.

We confirmed that we were saturating performance with the amount of training data, by varying the number of images we trained on, using values of 10, 50, 90, and 99 images per object (out of 100 total

images per object). Moreover, in a pilot experiment we established that we were not underestimating deep network performance due to the fact that the marmosets, humans, and macaques were able to generate an expanded "training set" of retinal activations because they were free to fixate anywhere during image presentation. To test this possibility, we generated an augmented training set that was 100x bigger, taking 100 random crops of different sizes and center locations (but not horizontal reflections) to mimic the varied fixations that marmosets and humans were allowed. We then trained classifiers on AlexNet representations, as that was the network with the lowest performance and thus greatest potential for improvement. It seemed plausible that this augmented dataset would lead to improved performance, as convolutional networks are not invariant to modest translation or scaling[58], and this kind of data augmentation is a central component of contemporary deep network training pipelines. Nonetheless, expanding the dataset hardly improved decoder performance at all, demonstrating that varied "fixations" do not substantially improve classifier performance on top of a pretrained network, and thus we appear not to be underestimating the difficulty of this task for deep network representations.

Classifiers for i1ns. To evaluate the i1n of artificial visual systems (deep neural networks and pixel representations), we computed the distance of each image from the hyperplane learned by a linear SVM. We took 50-50 splits of images, trained classifiers on each half and evaluated distances for the left-out images. To derive an i1n for each artificial visual system, we averaged the distance for each image over its value in the three tasks, and subtracted off mean values for each task. To get highly reliable estimates of the network's i1n, we performed this procedure 1,000 times for each task and each network, as well as for pixel representations. The ability to run arbitrarily large number of classifiers *in silico* results in the ability to drive the split-half correlation of the resulting i1ns arbitrarily high. Indeed, while the distances from the hyperplane are relatively reliable across individual partitions (for network classifiers, correlation coefficient ranges from 0.86 to 0.93; for pixels-based classifier: 0.66), the reliability of the distances averaged across 500 random train-test partitions is greater than 0.999 for all network classifiers and 0.998 for pixel-based classifiers. Because of these exceedingly high test-retest reliabilities, we did not apply noise correction to the classifier i1ns learned from pixel or deep network features.

Classifiers for rodent task. To evaluate the kinds of visual representations required to generalize on the task used by Zoccolan and colleagues, we mimicked the training that the animals received in how we trained our classifiers. We trained classifiers on pixels from the 28 images used initially in training (14 images at either 0° rotation or 1.0x scale for each of two objects), and evaluated the performance of the resulting classifier on the 80 held-out test images (40 images for each of the two objects). We again used a linear SVM classifier with a hinge loss and L2 regularization and selected the regularization strength hyperparameter via cross-validation within the 28 train images. We evaluated 5 candidate values for regularization strength which were logarithmically spaced, and performed 10 random splits within the

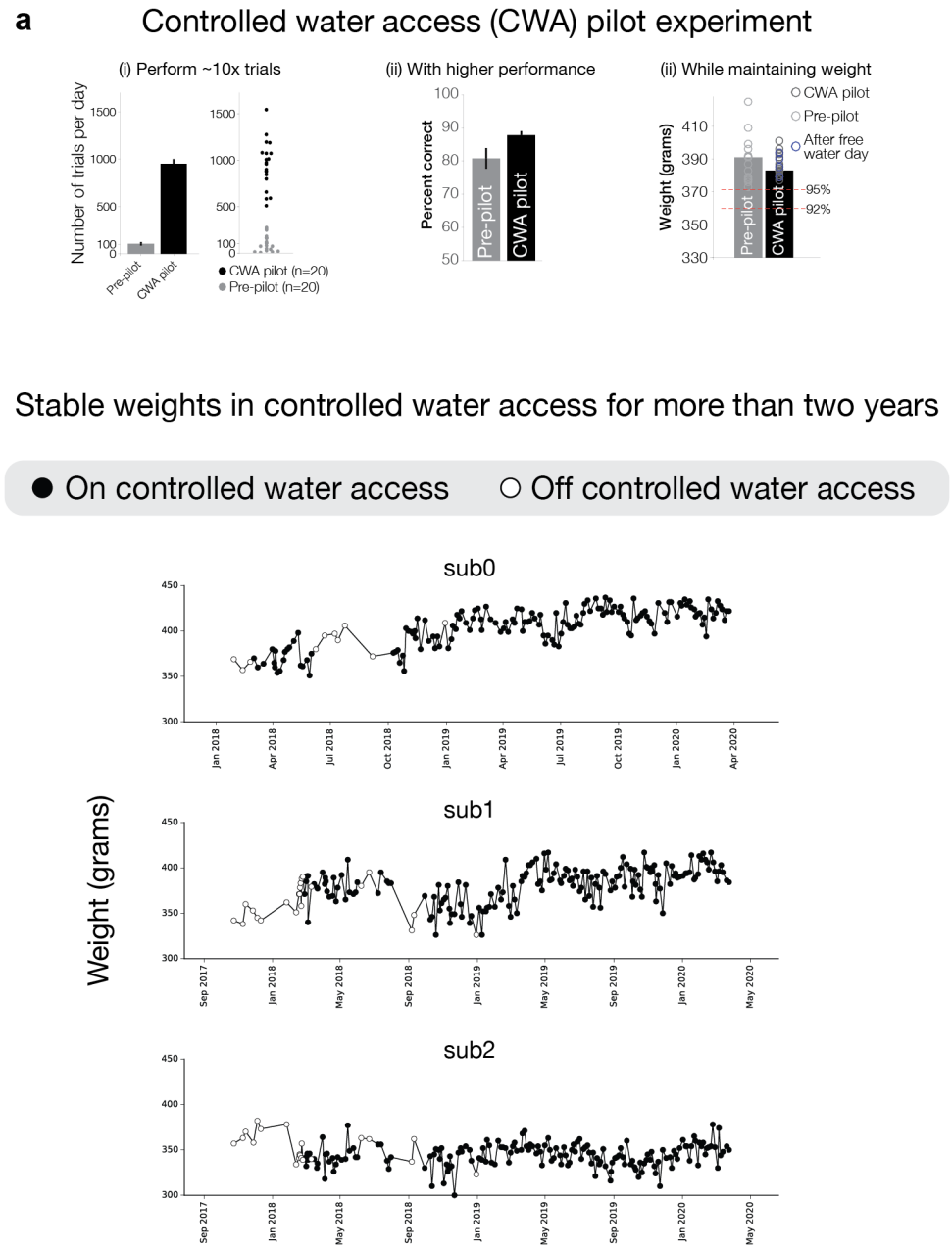
training set, training classifiers with each regularization strength on 23 images and evaluating the quality of the fit with the remaining 5. We then selected the regularization strength that performed best on left-out train images, and trained a single classifier with all 28 images with this regularization coefficient. We evaluated the performance of this classifier on the unseen 80 test images, and found that it classified all but two correctly (overall performance: 97.5%)—the two that it got incorrect were at the smallest size (0.5x) combined with the most dramatic rotation (60° or -60°) (bottom left and right corners of Fig. 3a). Given the already high performance of image pixels, we did not further evaluate deep network performance on this task (but see Vinken and Op de Beeck[30]).

Supplemental Figure 1. Species comparison table.

	Non-primate			Prosimian	Simian primates		
	Mouse	Rat	Treeshrew	Mouse lemur	Marmoset	Macaque	Human
Mass (kg)	0.025	0.400	0.160	0.060	0.300	7.500	80
Gyrification index	1.02[2,3]	1.04[2,3]	1.04[3]	1.11[2]	1.18[2,3]	1.75[2]	2.56[2]
Brain weight (g)	0.4	2	3	1.7	8	80	1500
Number of cortical areas	~40[4]	?	~40[5]	~50[6]	~120[4]	~140[4]	~180[4]
Visual acuity (cycles per degree)	0.5[7]	1[7]	2[8]	5[8]	30[8]	50[8]	60[8]
Peak retinal cone density (cones/mm²)	11,000-18,000[8]	?	36,000[8]	7,500-8,000[8]	152,000-360,000[8]	134,000-160,000[8]	100,000-324,000[8]
Eye ball size (axial diameter, mm)	5.2[9]	5.6[9]	8[9]	9.2[9]	11.3[9]	20[9]	24[9]
Diurnal or noctural	Noctural	Noctural	Diurnal	Noctural	Diurnal	Diurnal	Diurnal
Last common ancestor with humans (MYA)	>90[1]	>90[1]	>90[1]	87[1]	43[1]	32[1]	N/A

Supplemental Figure 1. Species comparison table.
The marmoset may offer the strengths of small animal models (small size, relatively flat brain), while preserving many desirable aspects of a larger primate model like the macaque (number of cortical areas and high acuity vision, which may stem from their evolutionary distance and visual ecology). Each column is a species, and each row a property of the animal. Arguably more desirable traits in a neuroscience animal model are shaded green, less desirable brown. MYA: Million years ago; gyrification index: the ratio of total cortical surface area over exposed cortical surface area. Estimates of the number of cortical areas vary depending on the method (cytoarchitecture, connectivity, etc.); however, one generally accepted trend is that the number of areas increases substantially in the simian primates[35]. Estimates of mouse and human number of areas are considered the most accurate because of the convergence across multiple measurements[36], while the number of cortical areas in emerging models such as the tree shrew and mouse lemur are less well established (though see[59,60]). 1: Perelman et al., 2011[33]; 2: Zilles et al., 2013[61]; 3: Ventura-Antunes et al., 2013[62]; 4: van Essen et al., 2019[36]; 5: Wong & Kaas, 2009[59]; 6: Saraf et al., 2019[60]; 7: Prusky et al., 2000[29]; 8: Kirk and Kay, 2004[21]; 9: Veilleux and Kirk, 2014[63].

Supplemental Figure 2. Controlled water access is effective and healthy in marmosets over extended periods of time.



Supplemental Figure 2. Controlled water access is safe and effective in marmosets over extended periods of time.

a. **Controlled water access pilot experiment.** In controlled water access paradigms, animals only receive fluid from performing the task. This approach, though standard in macaque and rodent work, had been largely avoided in marmosets, potentially due to a reputation of being too fragile. We evaluated the potential of using water restriction in marmosets, proceeding carefully at first with a small pilot experiment designed to test whether controlled water access may be safe and effective. We directly compared behavior of one animal when measured under controlled water access versus *ad libitum* water access. In an effort to give the *ad libitum* condition the greatest chance of success,

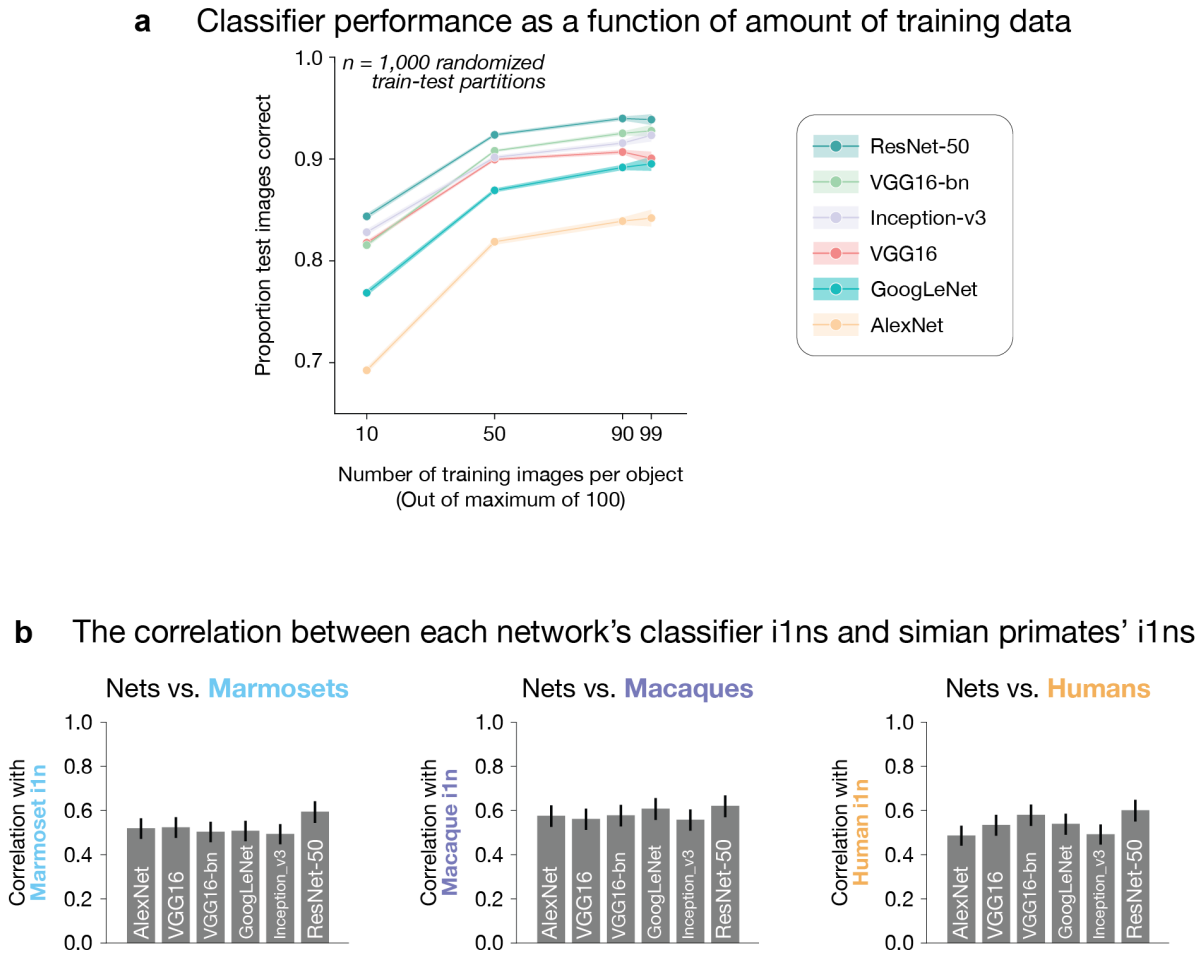
in this condition we removed the marmoset's access to water three hours before the task to induce a greater probability of thirst and used a high-value reward during the task (sweetened condensed milk). Nonetheless, under the controlled water access condition, the marmoset performed approximately 10x the number of trials (i), with higher performance (ii), while maintaining a relatively stable weight (iii).

b. **Stable weights in chronic controlled water access for more than a year.** In follow-up studies using more animals and longer time periods, animals maintained stable weights under controlled water access for more than a year.

Supplemental Figure 3. All 400 images.

Supplemental Figure 3. All 400 images. All four hundred images of camels, wrenches, rhinos, and legs superimposed on natural backgrounds. We measured marmoset and human performance on each of these images, and compared these performances with each other and with macaque performance on the same images collected in previous work[20].

Supplemental Figure 4. Deep networks: Task performance and i1ns.



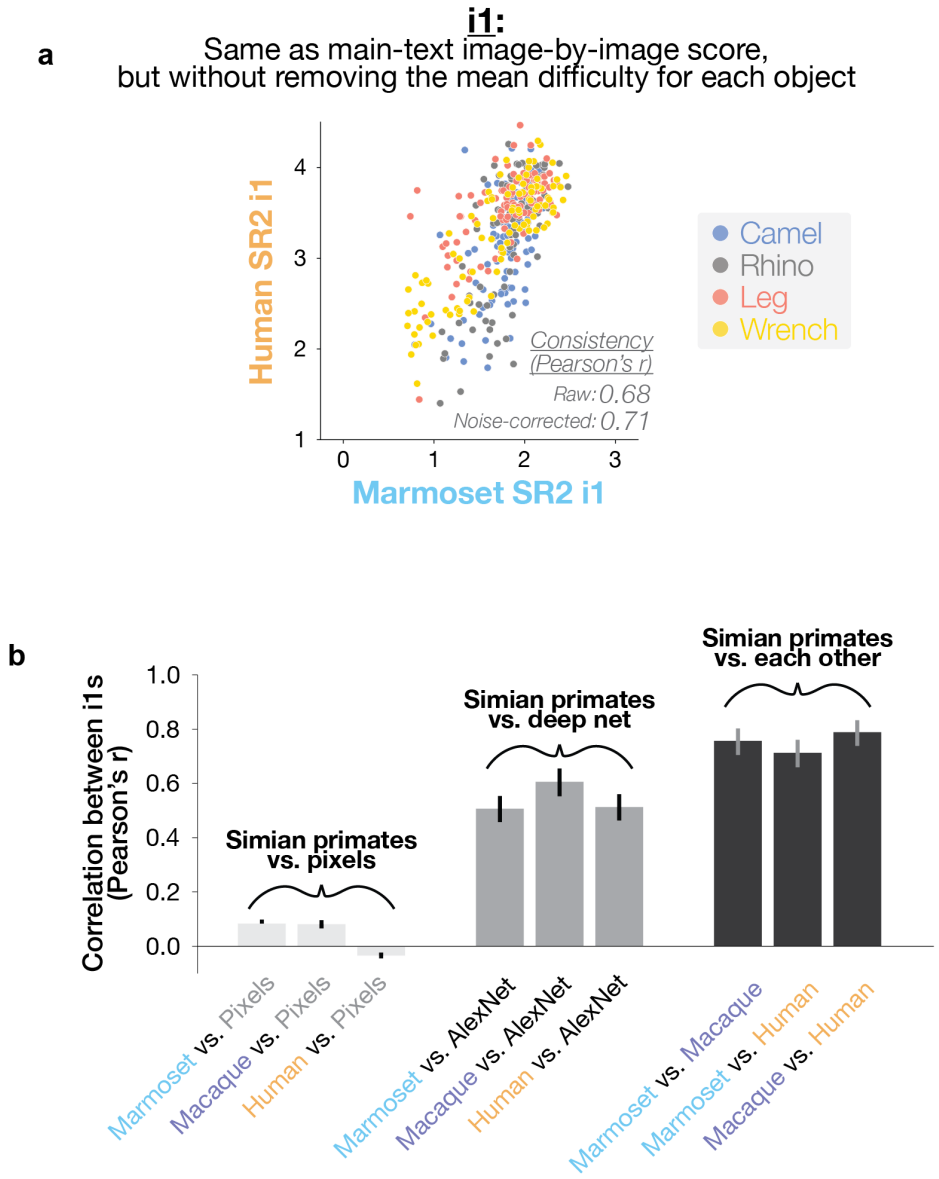
Supplemental Figure 4. Deep networks: Performance as a function of amount of training data and similarity to simian primates in i1n.

a. **Deep network performance on core recognition task.** The difficulty of a discrimination task varies drastically depending on the choice of stimuli[19], and so we verified that our images yielded a challenging task not only by assessing human performance as we report in the main text, but also by evaluating the performance of state-of-the-art engineering systems. We trained binary classifiers atop the penultimate layer of artificial deep neural networks, using 10, 50, 90, or 99 images per object (total images per object: 100), varying the training regime to establish that we reached a performance plateau with the amount of training data that we had available. Each line is a different network (VGG16-bn denotes a VGG16 architecture trained with batch normalization). Error bars are SEM over 1,000 classifiers trained on random train-test partitions; standard errors increase as amount of training data increases since the test dataset size concomitantly decreases (i.e., train size of 99 leaves only 1 image per object for testing). As we report in the main text, the raw input was insufficient to support task performance, as image pixel representations performed near the 50% chance level. Additional sensory processing, as instantiated by deep artificial neural networks, yielded performance at 84-94% indicating that high performance was achievable, but that even high-quality computer vision systems did not readily perform the task perfectly. These analyses complement the human

performance results in demonstrating that recognizing the objects in these images requires nontrivial sensory computation.

b. **Correlation between classifiers on different networks' features and simian primates**. From left to right, we compared marmoset, macaque, and human i1ns with i1n of the six networks. The consistency between networks and simians was relatively similar across the different networks.

Supplemental Figure 5. Image-by-image behavioral similarity between marmosets and humans is robust to details of the comparison metric.



Supplemental Figure 5. Image-by-image behavioral similarity between marmosets and humans is robust to details of the comparison metric.

a. The image-by-image difficulty score in the main text (i1n) had the mean difficulty of the object removed from each image-wise score (see main text Fig. 2a). Previous work had found that such a metric of within-object image difficulty is highly discriminating between model systems (Rajalingham et al., 2018). Here, we plot the same image-by-image metric, but without this object-level de-meaning step. We find that that the similarity between marmosets and humans is robust to this detail of the comparison metric. Additionally, as this metric does not include a de-meaning step, this plot shows the difference in overall performance level between marmosets and humans—marmoset image-wise d' ranges from just under 1 to approximately 2.5, while human d' scores range from just under 2 to just over 4. While overall performance varies across simians (as shown in main text Fig. 1e), the

comparative difficulty of each image was relatively robust as seen in the high correlation between marmosets and humans (here and main text, Fig. 2).

b. When we compare the unnormalized i1s from pixel classifiers, deep network classifiers, and simian primates, we see a similar pattern of results as we see with i1ns (compare to main text, Fig. 2g).