Object categorization based on abstract concepts by monkeys, humans, and vision models

Han Zhang[1, 2], Zhihao Zheng[1], Jiaqi Hu[1, 2], Qiao Wang[1], Mengya Xu[1], Zhaojiayi Zhou[1], Zixuan Li[1, 2], and Gouki Okazawa[1, 2, *]

[1]Institute of Neuroscience, Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China
[2]University of Chinese Academy of Sciences, China
[*]Correspondence: okazawa@ion.ac.cn

Abstract

Viewing an object in the external world evokes a rich variety of concepts associated with it in the human mind, such as whether it is animate or not, man-made or not, big or small, and countless other attributes. This concept recognition seems to depend on the human-specific knowledge about these objects. However, studies have shown that neurons in visual areas of non-human primates encode some object concepts, raising questions about the extent to which object concepts are uniquely possessed by humans. Here, we show that macaque monkeys can rapidly learn to classify natural object images according to a surprisingly rich set of rules defined by what humans would call abstract concepts. We tested more than 10 binary classification tasks, including animate versus inanimate, natural versus man-made objects, and mammalian versus non-mammalian animals. The monkeys learned each rule in a few days, generalized the learned rules to new images of objects they had probably never seen before and made error patterns consistent with human judgments. Visual deep neural networks (DNNs) could also perform the tasks that monkeys could learn, whereas both DNNs and monkeys failed to classify some less common abstract concepts. Thus, the monkeys' successful classification likely depended on whether concepts were embedded in natural images as visual features that could be extracted by high-level visual processing. The observed capacity of primate brains may subserve the formation of rich abstract concepts that the human mind possesses.

Introduction

For humans, objects in the world are imbued with concepts. We can immediately tell whether a visually perceived object is animate or inanimate, natural or man-made, edible or not, mammal or not, etc. The mechanisms by which humans acquire and recognize object concepts from visual images have been a central concern in the fields of cognitive science, neuroscience, psychology, ethology, anthropology, philosophy, and linguistics. Earlier theories proposed that humans have preferential access to an intermediate level of abstraction (basic-level categories, such as dog and cat) (Rosch et al. 1976) because their exemplars are likely to have similarities in their visual appearance, functions, and other attributes. More abstract categories, such as "animal" (i.e., superordinate categories; here we refer to them as "abstract object concepts") are considered to have fewer common, defining features (Rosch et al. 1976). Many theories propose that the formation of such abstract concepts depends on the knowledge related to objects (Murphy and Medin 1985; Ralph et al. 2017), in which experience and human language play important roles (Murphy 2004; Vong et al. 2024).

Visual neuroscience has also long investigated neural representations of objects at various levels of abstraction, with significant contributions from studies using macaque monkeys. In the human brain, neuroimaging and neuropsychological studies have revealed gradients in the representation of object categories in higher visual cortices (Clarke and Tyler 2015; Contier et al. 2024; Grill-Spector and Weiner 2014; Mahon and Caramazza 2009, 2011; Martin 2007) or even in extra-visual areas (Huth et al. 2012), where there are regions that preferentially encode animals, tools, places, faces, bodies, food, and other kinds of categories (Clarke and Tyler 2015; Mahon and Caramazza 2009, 2011; Martin 2007). Many of these have also been found in the macaque brain (Hesse and Tsao 2020; Vogels 2022), and importantly, population neural responses of the monkey higher visual area clearly reflect the animate vs. inanimate boundary, just as the human higher visual area does (Bao et al. 2020; Kiani et al. 2007; Kriegeskorte et al. 2008; Liu et al. 2013; Zhang et al. 2024). Recently, studies have demonstrated that deep neural networks (DNNs) effectively predict neural responses to natural images in monkey visual areas (Kar and DiCarlo 2024; Yamins et al. 2014), with similar predictive success observed for human visual areas by using neuroimaging (Khaligh-Razavi and Kriegeskorte 2014). DNNs also partially account for abstract concept encoding, such as animate vs. inanimate concepts(Jozwik et al. 2017; Peterson et al. 2018).

While all these extensive works seem to support the similarity of human and monkey object vision (Koopman et al. 2017; Rajalingham et al. 2015), whether monkeys can truly classify objects at abstract levels as humans do remains underexplored. They may not have this classification ability if knowledge and language play a role, but the aforementioned physiological studies have shown the similarity of object encoding in the brains of the two species. A series of seminal works by Fabre-Thorpe et al. (Delorme et al. 2000; Fabre-Thorpe 2003; Fabre-Thorpe et al. 1998; Fize et al. 2011) indeed revealed that monkeys can report the presence or absence of an animal from rapidly presented natural photographs using a Go-NoGo paradigm. Several other notable works have shown that monkeys can classify trees vs. non-trees (Vogels 1999), humans vs. non-humans (Schrier and Brady 1987), monkeys vs. non-monkeys (Yoshikubo 1985), food vs. non-food (Fabre-Thorpe et al. 1998; Santos et al. 2001), cats vs. dogs (Eldridge et al. 2018; Minamimoto et al.

2010), and cars vs. trucks (Minamimoto et al. 2010). However, each study typically tested one or, at most, two classification rules, many of which were basic-level categories whose exemplars had similar appearances. Furthermore, studies often used Go-NoGo detection paradigms (Eldridge et al. 2018; Fabre-Thorpe et al. 1998; Minamimoto et al. 2010), which are not typical for assessing object classification in humans. Therefore, a systematic investigation of monkey behavior is much needed, using a categorization task with various abstract rules to assess the similarity of abstract concept classification between monkeys and humans. To compare monkey ability across rules, each rule must be trained and tested in a short, fixed period of time so that the monkey behavior is not biased by overtraining.

Here, we develop a behavioral paradigm to overcome these challenges and show that macaque monkeys can rapidly learn to classify object images into a surprisingly rich variety of abstract concepts using a battery of binary categorization tasks (more than 10 tasks, 26 including subtasks, $\sim$315K trials from three monkeys) together with human behavioral data in the same tasks ($\sim$98K trials). The monkeys learned classifications such as animate vs. inanimate, natural vs. artificial, mammal vs. non-mammal, big vs. small, within a few days of experimental sessions, even though they likely had no prior experience of interacting with many of the objects we presented. Their behavior was correlated with that of humans and could not be explained by responses to low- or mid-level image features. We wondered how monkeys could perform these tasks and found that visual DNNs could also perform these tasks despite not being explicitly taught about these concepts. Together, these results suggest that primate brains are equipped with the capacity to extract high-level visual features correlated with abstract object concepts and readily associate them with behavioral responses.

Results

Monkeys quickly learn many abstract concept classification tasks

Our task is a version of binary decision-making paradigms, in which monkeys categorize an object image into one of two options according to a hidden rule (Fig. 1a). The experiments were conducted using a touchscreen system installed on the monkey's home cage (Supplementary Fig. 1a). During each trial, after the monkey touched a red dot, an object image appeared together with two gray target boxes. The monkey had to touch this object image and move it to one of the two target boxes by dragging the image. The image position moved along with the monkey's finger until it reached a box. Upon reaching, the two boxes disappeared, and the same object image was revealed under the correct box, together with a juice reward for the correct choice. We reasoned that the direct interaction with an object image encouraged monkeys to rely on the properties of the object to make a behavioral choice. Furthermore, this dragging motion required monkeys to slow down their choice report (> 0.5 sec to reach a target), potentially suppressing random impulsive choices and allowing them to make deliberate decisions (Kaufman et al. 2015). The categorization rule remained the same within a session, and the same rule continued across sessions until the monkey's performance plateaued. The monkeys had to infer this rule based on the feedback they received on

each trial.

After the monkeys were trained on this task structure using amorphous shape stimuli (Supplementary Fig. 1b, c), we started to show naturalistic object stimuli with categorization rules defined by abstract concepts. Figures 1b-f show the results of the first experiment we conducted, in which the monkeys were trained to classify object images into animate or inanimate categories. The training images for each category consisted of a variety of exemplars (Fig. 1b; the animate category included humans, monkeys, mammals, and reptiles/amphibians, whereas the inanimate category included plants, food, tools/everyday objects, and electronics). In this first line of experiments, we used grayscale images with the background scene cropped to exclude color information.

To our surprise, all three tested monkeys achieved high classification performance in the animate vs. inanimate categorization task within a few days of training (averaging 760 trials in an 1-2 hour session per day), even though they had to infer the common rule from many different images (Fig. 1c-d). They started to show above-chance performance from Day 1 ($p < 0.011$ for all monkeys, binomial test), and their correct rate monotonically improved over the days, plateauing at approximately Days 4-6 with a performance of $\sim$90% (Fig. 1d). The same training images ($n = 96$) were repeatedly presented during these sessions, but the monkeys started to show above-chance performance only after a few repetitions per image (Fig. 1d inset; $p < 0.05$ at 3, 4, and 9 repetitions for monkeys Ol, Ju, and El, respectively, binomial test).

Although their overall performance did not reach a 100% correct rate in six days, we found that they achieved a nearly 100 % correct rate for some images and performed poorly for some other images. In Figure 1e, the correct rate for each image is plotted over the last four sessions, showing the diversity of correct rates across images that cannot be explained by chance ($\chi^2 > 715$, $p < 10^{-10}$ for all monkeys, χ^2-test). Interestingly, images with a nearly 100% correct rate were often prototypical examples of animate and inanimate categories such as monkeys, humans, and key locks. In contrast, images with correct rate less than 50% tended to be atypical examples, such as a snake in a coil or a roasted chicken (defined as inanimate food). The patterns of correct rates were also largely consistent across monkeys (average of three comparisons: $R = 0.49$, $p < 2.1 \times 10^{-4}$ for all comparisons). The reaction times (RTs) also tended to be longer for these stimuli (Fig. 1f), and the drag trajectories for these stimuli occasionally showed patterns that resembled those of the animals were changing their minds (Supplementary Fig. 2; Kaufman et al. 2015). In a later section, we empirically test the similarity of these behaviors to human judgments.

We then confirmed that the monkeys could generalize the rules that they learned to new images seen for the first time (Fig. 1i). After the learning sessions, we introduced a new set of images ($n = 96$) that largely overlapped basic-level categories with the learned images but had different appearances (Fig. 1b). The monkeys' correct rates for these new images were high from the first exposure ($p < 6.6 \times 10^{-5}$, binomial test). To further confirm that the results could not be accounted for as learning of the individual associations of stimuli and target locations, we performed randomized control tasks in which the two target locations were associated with random object images so that there was no common rule to classify them (Supplementary Fig. 3a-b). Although the monkeys could still learn to associate individual images with targets, their learning speed and performance were worse (Supplementary Fig. 3c-d).

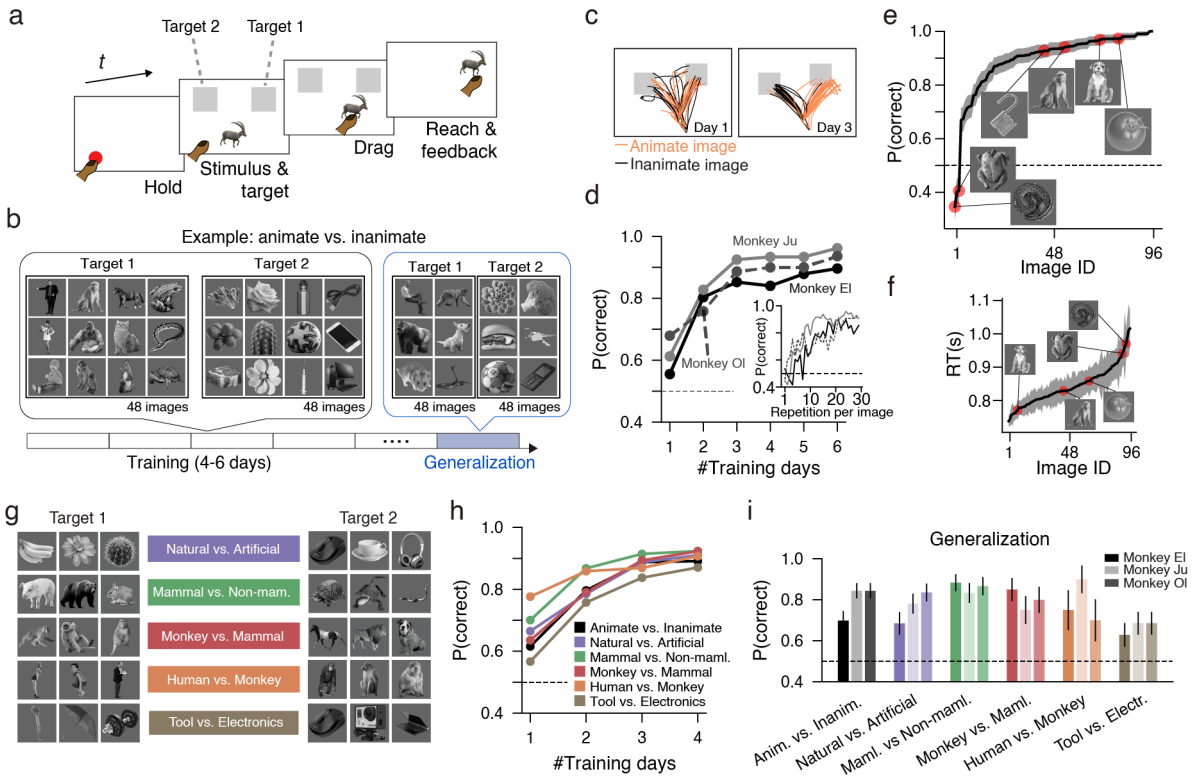

Figure 1: ***Monkeys rapidly learn various abstract object concept classification tasks.*** *(**a**) Object drag task. In each trial, after the monkey held a fixation dot, an object image and two gray target boxes appeared. The monkey had to touch and drag the image to the correct box to receive a reward. Two boxes were associated with different object concepts. (**b**) We used grayscale photographs of objects with no background in this first set of experiments. We used ∼100 training images to train monkeys and measured their generalization performance with test images. (**c**) Reach trajectories showed that monkeys tended to choose targets randomly on Day 1, but by Day 3, they chose correct targets in many trials. Examples from monkey Ju. (**d**) All monkeys' correct rates rapidly improved during the first three days of the animate vs. inanimate task, reaching 85-90% and then plateaued (chance level: 50%). Before the performance plateaued, monkeys saw each training image 10-20 times (inset). (**e**) After learning of the animate vs. inanimate task (post-Day 3), monkeys made more than 90% correct choices for many images, but performed poorly on certain images. The plot is the average of the three monkeys, with images ordered by correct rates. The shading indicates S.E.M. across trials. (**f**) Reaction times (RTs), measured as the time from stimulus onset to reaching a target, also varied across images. The plot is the average of the three monkeys, with images ordered by RTs (see also Supplementary Fig. 9b). The shading indicates S.E.M. across trials. (**g**) We tested a variety of classification rules in succession. Each task used 60-96 grayscale training images. (**h**) Monkeys learned all tasks with similar learning rates. The plot showed the average of the three monkeys. (**i**) After training on each task, the monkeys could generalize the rule to new stimuli they saw for the first time. The error bars represent S.E.M. across images. All images were used either under CC0 license or Pixabay content license.*

Encouraged by these results, we tested a variety of classification rules based on both superordinate and basic-level object concepts (Fig. 1g). These tasks were natural (plants, food) vs. artificial (man-made) objects, mammal vs. non-mammal animals (reptiles/amphibians), monkeys vs. mammals, humans vs. monkeys, and tools vs. electronics. The monkeys had no difficulty learning these tasks in a few days (Fig. 1h; $p < 10^{-10}$ for all monkeys in all tasks on Day 4, binomial test). As in the animate vs. inanimate task, the monkeys learned many images easily and showed poor performance for some other images (Supplementary Fig. 4). And they generalized the rule to new images that they had never seen before (Fig. 1i; $p < 6.1 \times 10^{-6}$ in all tasks aggregated across monkeys, binomial test). When we switched to the next rule, we did not give the monkeys any explicit cues, but they learned to adopt the new rule quickly.

Thus, monkeys rapidly learned to classify naturalistic objects based on rules defined by various abstract object concepts. In what follows, we show that their classification ability extends to a larger number of natural image sets, is similar to human behavior, and cannot be explained by the detection of low- or mid-level image features. We will then discuss how monkeys might have performed these tasks.

Low-level image features do not explain monkey behavior

After successfully training monkeys on various classification rules using small sets of controlled stimuli, we set out to collect behavioral data using a larger number of natural images ($>$ 3,000). One goal was to challenge monkeys' ability to use images with even more diverse appearances and natural backgrounds. The second goal was to rule out the possibility that the monkeys could generalize rules on the basis solely of the similarity to learned images (e.g., an exemplar-based strategy; see below). The third goal was to quantitatively inspect what the features monkeys used to solve the tasks using more diverse images.

We created tasks using images from the THINGS database (Hebart et al. 2019; Stoinski et al. 2024), a large-scale, human-curated object image database (Fig. 2a). The database contains 1,854 concrete object concepts (e.g., dog, spider, lemon, and table; referred to here as "object categories" to avoid confusion) that are frequently used in the American English language and include natural images associated with each category. We selected a subset of the object categories to create three tasks (Fig. 2a): animate vs. inanimate (373 object categories, 3,683 images in total), natural vs. artificial objects (230 object categories, 1,915 images in total), and mammal vs non-mammal animals (143 object categories, 1,668 images in total). For each task, we selected 100 images as a training set, trained the monkeys until their performance plateaued (85-95%), and then introduced the remaining images to test generalization. Importantly, the test images, which were shown only once, could be from object categories that were shown during training ("old"; Fig. 2a, second row) or were never shown ("new"; Fig. 2a, bottom two rows). This enabled us to test whether monkey performance depended on the similarity to the trained images (i.e., whether the "old" categories showed greater performance).

The monkeys performed these tasks without difficulty, despite experiencing much greater vari-

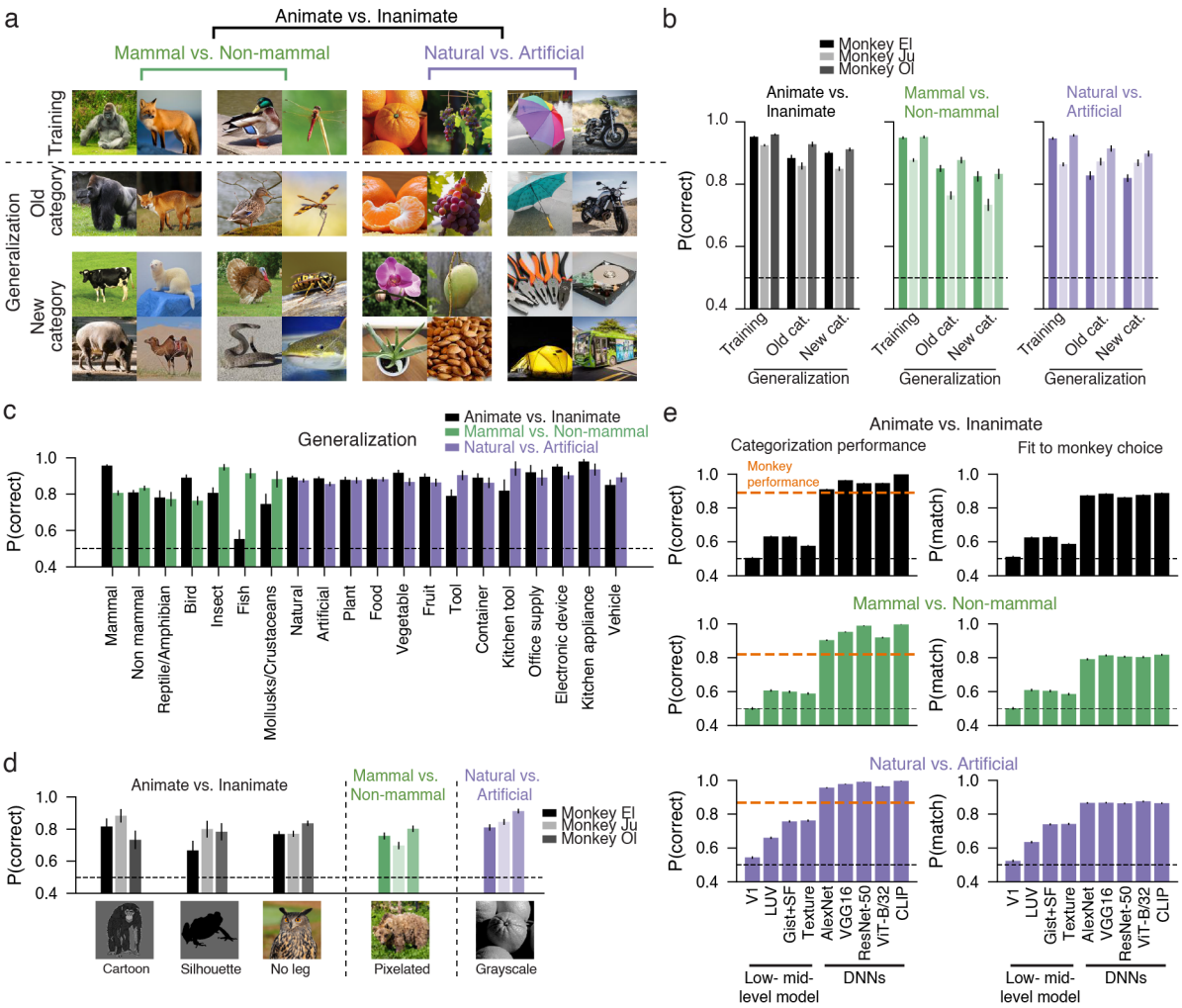

Figure 2: ***Behavioral performance for large-scale image sets cannot be explained by low- or mid-level visual features.*** *(**a**) We assessed monkey performance in three tasks using large-scale image sets from the THINGS database (Hebart et al. 2019). We used 100 images to train monkeys (top row) and then presented 1,500-3,600 generalization images, containing "old" object categories present in training images (second row) and "new" categories that were never shown (bottom two rows). Each task used different training images. (**b**) All monkeys performed well on generalization images (75-93%) for both the "old" and "new" object categories. (**c**) Performance was generally high across a range of object categories (see Methods for the classification of images). (**d**) Monkeys also performed well on control images that lacked certain visual features, such as cartoons without naturalistic textures, silhouettes without internal structures such as faces, and grayscale images without color. (**e**) Compared with deep neural networks (DNNs), models of low- and mid-level visual features cannot perform classification (left) and fit monkey behavior (right) well. We constructed a linear classifier of the correct choices or monkey choices from the output of each model using the training images and plotted their performance on the generalization images. $P(match)$ indicates the probability that the monkey and model choices matched. The orange dashed lines represent the classification performance averaged over the three monkeys. (**b-e**) Error bars indicate S.E.M. across object categories. All images were used either under CC0 license or Pixabay content license.*

ability in image appearance than that in the previous image sets (Fig. 2b; Movies 1-3; $p < 10^{-10}$ for all monkeys in the three tasks, binomial test). Importantly, their generalization performances to "new" object categories were largely comparable to that of the novel images of the "old" object categories they learned during training (Fig. 2b; $U = 2,360 - 14,902$, $p = 0.086 - 0.98$ across monkeys and tasks, two-tailed Mann-Whitney U-test, $\mathrm{BF}_{01} > 3.7$ except for the mammal vs non-mammal in monkey Ol whose BF_{01} was 1.9). Their performances were also largely similar across a range of object categories in the datasets (Fig. 2c), except for "fish" images during the animate vs. inanimate task. Thus, it is unlikely that the monkeys generalized the rules by judging the similarity to specific exemplars they learned during the training.

We then wondered whether the monkeys relied on a particular feature in the images to perform the tasks (e.g., face, leg-like shape, certain colors, textures), so we presented control images that lacked these features (Fig. 2d). After the monkeys learned the animate vs. inanimate task, we presented cartoon images of animals and objects that lacked naturalistic textures, and silhouette images that retained only the outlines of objects without internal structures such as faces. In the THINGS image set, there were close-up images of animals without legs. The monkeys performed well on these images (Fig. 2d; 73-88 % correct, $p < 6.7 \times 10^{-3}$ for all monkeys, binomial test). In contrast, they could not generalize the rule to "texform" images (Kramer et al. 2023; Long et al. 2018), which retained only mid-level image features (Supplementary Fig. 5). In the natural vs. artificial task, color could have been strong cues (natural objects tend to be greenish); therefore, we presented a grayscale version of some images and confirmed that the monkeys performed well on these images as well (81-91 % correct, $p < 10^{-10}$ for all monkeys, binomial test). Finally, monkeys could have relied on fur texture to perform the mammal vs. non-mammal task, but the monkeys performed the task well even after distorting texture (distortion applied using a Photoshop filter, see Methods; 70-80 % correct, $p < 10^{-10}$ for all monkeys, binomial test). Thus, they likely did not rely on a single specific feature to solve the tasks.

We further confirmed that vision models based on lower-level image features could not explain monkey behavior well (Fig. 2e). These models included V1-level filter outputs (Pinto et al. 2008), luminance- and color-based statistics, spatial-frequency summary statistics, and texture statistics (Portilla and Simoncelli 2000) derived from correlation structures of V1-level outputs. For all the models, we first reduced their output dimensions using principal component analysis (PCA), constructed a linear classifier for either the true categories of the training images in a task or the monkey's choices for the training images, and then tested their performance on the generalization set (see Methods for details). The classification performance of the low- and mid-level models was typically $\sim$60%, and they were all significantly lower than the monkey performance ($p < 10^{-10}$ for all comparisons between models and average monkey performance, binomial test, Bonferroni corrected for multiple comparisons). Correspondingly, even when these models were directly fit to monkey behavior, the achieved rates of choice agreement between the models and the monkeys were low ($\sim$60% in many cases). These results did not depend on specific details in the dimension reduction and regression methods (Supplementary Fig. 6).

In contrast, we found that various deep neural networks (DNNs) could perform the tasks and obtained a better fit to monkey behavior. We tested several model architectures, including AlexNet (Krizhevsky et al. 2012), VGG16 (Simonyan and Zisserman 2014), ResNet-50 (He et al. 2015), and

Vision Transformer (ViT-B/32) (Dosovitskiy et al. 2020), which were pretrained on ImageNet. We also tested CLIP (Radford et al. 2021), a recent large-scale model that was trained to match visual and text inputs, thus providing additional language supervision. Irrespective of architecture and learning objectives, all DNNs performed our tasks with high accuracy (>90%). Furthermore, these DNNs fit the monkey behavior better than the low- and mid-level visual models did($p < 10^{-10}$ for all comparisons between models, binomial test, Bonferroni corrected for multiple comparisons). In these models, deeper layers tended to better explain the monkey behavior (Supplementary Fig. 7). Better fits by DNNs and poor performance by other models were also consistently observed for the cropped, grayscale image sets tested in Fig. 1 (Supplementary Fig. 8a).

Overall, these results indicate that monkeys were unlikely to rely on the similarity to specific exemplars to generalize the classification rule (Fig. 2b) and that low-level image features could not account for their performance (Fig. 2d-e). In contrast, DNNs could solve these tasks and provide a better explanation of monkey behavior. This is not a proof that DNNs and primate brains share underlying computations; we have shown their correlations only in a limited stimulus space and will likely find their differences when we use even more diverse or controlled image sets and tasks (Bowers et al. 2023). Nevertheless, the results demonstrate that these abstract concepts can be linearly classified by sufficiently complex neural networks trained to perform generic object classification (Jozwik et al. 2017; Kubilius et al. 2016; Peterson et al. 2018). Thus, it would be feasible to interpret that monkeys could perform these tasks through high-level visual processing even without direct episodic experience with many of the objects we presented. We elaborate on this interpretation in the final section of Results.

Monkey categorization was moderately correlated with human behavior

To test whether the patterns of errors made by the monkeys for individual images (Fig. 1e) were similar to human judgments, we collected human behavioral data using the same experimental design with grayscale image sets (Fig. 1b, g). As in the monkey experiment, human participants had to infer the rule based on feedback alone. Nonetheless, they understood the rules within tens of trials and showed near-perfect performance thereafter in all tasks (Fig. 3a). This indicates the obvious superiority of humans in recognizing and adopting the conceptual rules over monkeys, but importantly, human participants still showed different reaction times (RTs) across images in each task (Fig. 3b, $F(95, 570) = 2.88$, $p < 10^{-10}$ for animate vs inanimate, $F > 2.14$, $p < 4 \times 10^{-6}$ for the other tasks, repeated-measures one-way ANOVA). Thus, we expected that difficult stimuli would cause longer RTs in humans and lower choice accuracy (and longer RTs; Fig. 1f) in monkeys.

We therefore constructed a metric of stimulus difficulty that embraces both choice accuracy and RTs to compare them between two species. We used the drift-diffusion model (Fig. 3c), which has been widely used to explain choices and RTs in various experimental settings, including decision-making tasks involving object recognition (Carlson et al. 2014; Heidari-Gorji et al. 2021; Luo et al. 2025; Mack and Palmeri 2011). The model assumes a one-dimensional stochastic internal state that meanders between two decision bounds corresponding to two choice options until it reaches

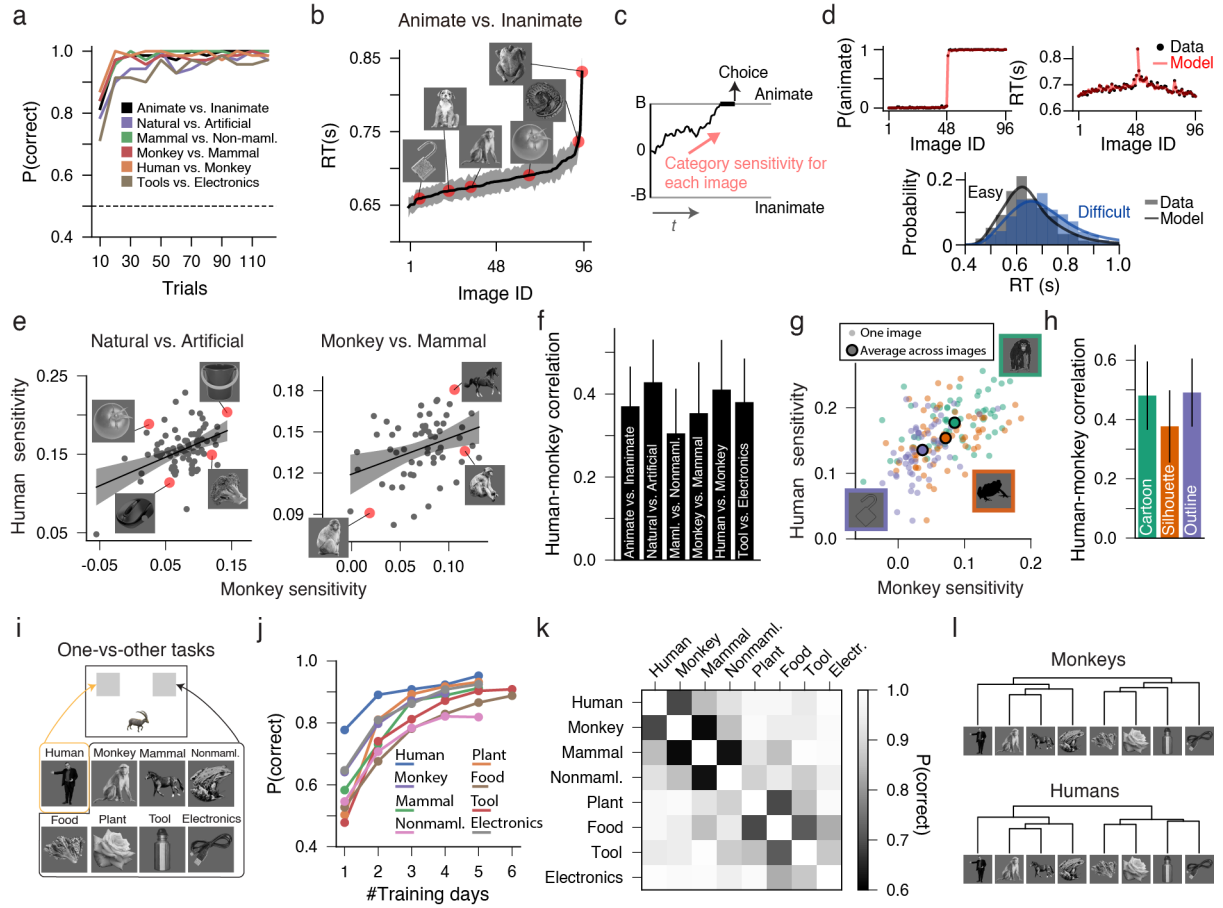

Figure 3: ***Monkey behavioral performance was correlated with human judgment.*** *(**a**) Humans quickly learned the tasks in Fig. 1 within 20-30 trials. The plots are the average of seven participants per task. (**b**) Different average reaction times (RTs) across images, averaged across participants. (**c**) We fitted the drift-diffusion model (DDM) to the choices and RTs. The model makes a decision based on a stochastic state biased by the category sensitivity of each image. (**d**) DDM accurately fits choices, RTs, and RT distributions. The plots are from the animate vs. inanimate task. Image IDs were sorted by category sensitivity. RT distributions were made using the trials aggregated across the top (easy) and bottom (difficult) 10 images. See Supplementary Fig. 9 for all tasks. (**e**) Example comparison of category sensitivity between humans and monkeys. Each dot represents an image, whose absolute category sensitivity was averaged across participants. The line is a linear regression, and shading indicates the standard error of the prediction line. (**f**) Positive correlations between humans and monkeys. Error bars indicate S.E.M. across images. (**g**) The category sensitivity for control images —cartoon, silhouette, and outline— was in the same order between monkeys and humans. (**h**) Positive correlations were also evident within each control image set. (**i**) To examine the category hierarchy, we trained a monkey to perform one-versus-other task for eight object groups (see text). (**j**) The monkey learned each of the eight one-versus-other tasks with similar learning curves. (**k**) The monkey's confusion matrix created from the performances on the eight tasks showed a clear tendency to confuse some pairs. (**l**) A dendrogram generated from the monkey confusion matrix (top) shows a hierarchical categorical structure similar to that of the human dendrogram (bottom; see Methods, Supplementary Fig. 11). All images were used either under CC0 license or Pixabay content license.*

one of them (Fig. 3c). An easy stimulus causes a large drift (higher sensitivity) toward the correct bound, leading to higher accuracy and shorter RTs. The slope of the drift is termed "category sensitivity" here, whose absolute value reflects how easily an image can be categorized correctly. Note that we did not assume that an evidence accumulation process underlies our task (Stine et al. 2020). Rather, we used the model as a means of deriving a single metric of stimulus difficulty (the absolute value of category sensitivity) for each image from the choices and RTs. Critically, the model fits both human and monkey data well (Fig. 3d, Supplementary Fig. 9; $R = 0.99 - 1.0$ for choices and $R = 0.81 - 0.99$ for RTs across monkeys, humans, and tasks), demonstrating its utility.

For all tasks, the absolute value of the category sensitivity of humans and monkeys across images had moderate positive correlations (Fig. 3e, f). Humans obviously had higher category sensitivity for all images, as expected from their near-perfect performance. Nonetheless, there was a tendency for images with higher category sensitivity for humans to also be higher for monkeys (Fig. 3e). The correlations of sensitivities between humans and monkeys were significantly positive ($R = 0.30 - 0.43$; $p = 7.4 \times 10^{-5} - 6.0 \times 10^{-3}$) for all tasks (Fig. 3f). We confirmed that similar results can also be obtained even when we use a different metric without model fit (Supplementary Fig. 10). As a reference, we compared the sensitivity between human participants randomly split in half and found that human-human correlations were greater ($R = 0.49 - 0.71$). This could be either because the monkey data were noisier or there were inherent, yet uncharacterized, differences in how human and monkey categorized images.

We also checked the two species' responses to pictures outside the range of natural images and again found the similarity between them (Fig. 3g, h). In the animate vs. inanimate task, we showed cartoon, silhouette, and outline images (Fig. 3g; part of the results are shown in Fig. 2d). The monkeys showed higher performance for cartoon and silhouette images than for outline images, which was consistent with slower human RTs for outline images. We also found significant positive correlations of category sensitivities between humans and monkeys within each control set (Fig. 3h). Note that DNNs are also known to have poorer classification performance for outline images than for silhouette images, even though they must retain the same shape information (Supplementary Fig. 8b; Baker et al. 2018).

Using another variant of the task paradigm, we were even able to compare the hierarchical structure of monkey confusion in object classification with human judgment (Fig. 3i-l). We selected eight major object groups (humans, monkeys, non-primate mammals, non-mammal, food, plants, tools/everyday objects, electronics) and attempted to quantify the perceived similarity of individual pairs with one monkey. For each object group, we performed a one-versus-other task, in which the monkey had to associate one target with the selected group and the other target with the rest (Fig. 3i). Thus, eight tasks were performed in total. The monkey showed a stereotypical learning curve for all tasks, reaching a plateau at 3-4 days (Fig. 3j), but tended to confuse certain pairs even after learning; for example, in the food-versus-other task, the monkey often moved plant images to the target associated with food. From the confusion patterns, we generated a dissimilarity matrix of the eight groups and found that the monkey indeed tended to confuse groups that appeared conceptually proximal to humans (Fig. 3k; see Methods section and Supplementary Fig. 11a for thematrix generation process). This pattern was very similar to the human similarity judg-

ments obtained in a multi-object arrangement task (Supplementary Fig. 11b; Kendall $\tau = 0.70$, $p < 10^{-10}$). The dendrograms generated from the human and monkey data revealed a similar hierarchical structure of the eight groups (Fig. 3l).

Object concepts classifiable by monkeys are embedded as image features

While monkeys could learn and generalize all the categorization rules we have tested thus far, it must be impossible for monkeys to learn every abstract concept that humans possess. Identifying what they can and cannot learn would be informative for understanding how they solve these tasks. We therefore created five additional conceptual categorization tasks to challenge the monkeys' ability (Fig. 4a). As often investigated in the vision science literature, we first tested outdoor vs. indoor scenes (Zhou et al. 2017) and big vs. small object categorization (Konkle and Oliva 2012; Long et al. 2016). Big vs. small objects was determined on the basis of physical object size (objects bigger than a chair were classified as "big"; Konkle and Oliva 2012) and had no correlation with image size. We had two versions of big vs. small, one using the set from Konkle et al. (2012) and one using images from the THINGS database. We then created two more tasks that would rely heavily on human knowledge and were unlikely to have common visual features to discriminate. These categories were "fire-related" vs. "water-related" objects such as lighter, fire extinguishers, faucets, and bathtubs, etc., and "Western culture" vs. "Eastern culture" objects such as mooncakes, Chinese knots, crowns, and Easter eggs, etc. For each task, we first trained the monkeys with training images for 4-6 days and then examined their generalization performance with test images as in Fig. 1b.

The monkeys successfully generalized rules in some tasks and failed in others, and visual DNNs exhibited similar patterns (Fig. 4b). The monkeys successfully generalized the outdoor vs. indoor rule to new stimuli ($p < 4.3 \times 10^{-7}$ for all monkeys, binomial test) and also performed well on the two versions of the big vs. small task ($p < 2.4 \times 10^{-6}$ for all monkeys in both tasks, binomial test), as did visual DNNs such as ResNet-50. Recognizing object size seems to require prior knowledge about these objects, but previous studies have already reported that this categorization is partially possible even based on mid-level textural visual features (Kramer et al. 2023; Long et al. 2016, 2018). In contrast, the monkey performance was substantially lower for the fire- vs. water-related task (46-64% correct). And their performances were at chance level for the Western vs. Eastern object task ($p = 0.38 - 0.86$, $\mathrm{BF}_{01} = 4.9 - 7.4$ across monkeys, binomial test). Notably, visual DNNs such as ResNet-50 trained on the ImageNet also displayed poorer performance on the fire vs. water task (53-75 % correct across models) and on the Western vs. Eastern task (46-60 % correct across models; see Supplementary Fig. 8). The performance of a language-informed model (CLIP) was better and above chance in both tasks ($p < 6.9 \times 10^{-6}$, binomial test). The human participants performed these tasks almost perfectly, suggesting their superiority over monkeys and DNNs.

In line with the classification performance, we found that the abstract concepts classified by the monkeys were linearly embedded as a major axis in the output of the visual DNNs (Fig. 4c). We visualized the low-dimensional embedding of the DNN responses to the images in each task

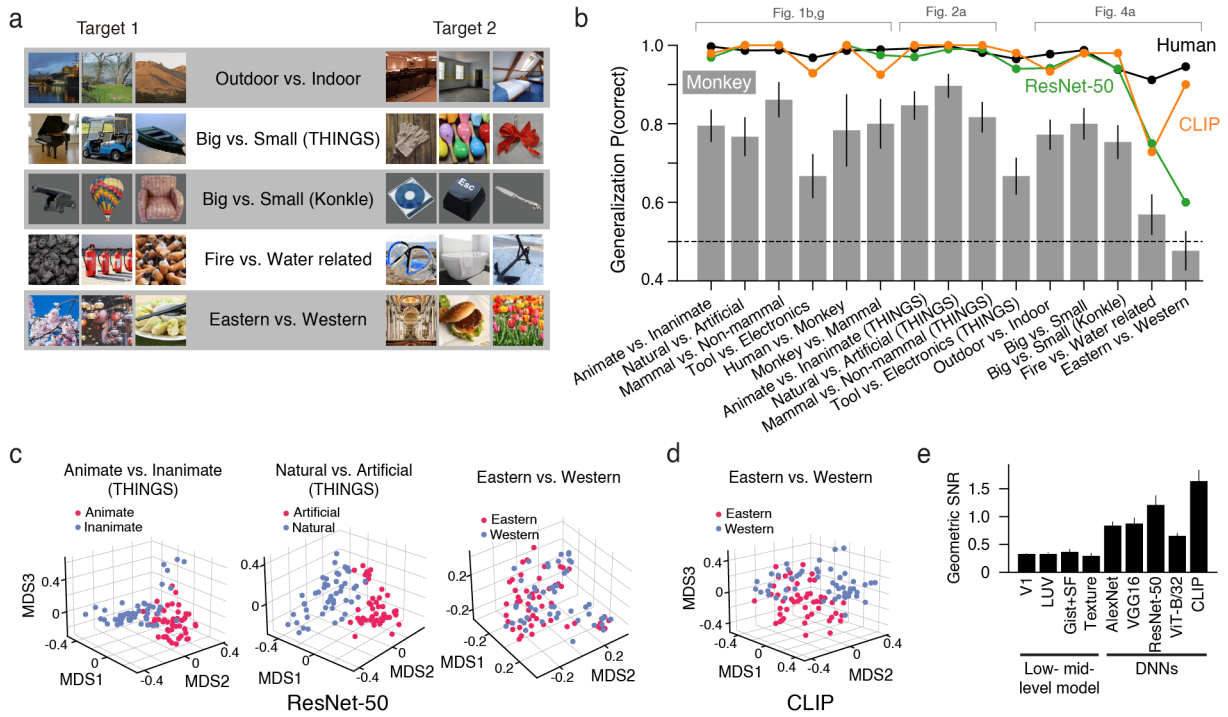

Figure 4: ***Object concepts classifiable by monkeys are embedded in image features.*** *(**a**) To determine the limit of monkeys' ability to classify concepts, we tested more concept tasks. As with the other tasks in Fig. 1, we trained the three monkeys with 90-120 images and assessed their generalization performance with 90-120 new images. (**b**) The monkeys ($n = 3$) performed well on tasks such as big vs. small objects and indoor vs. outdoor scenes, but they failed on some abstract tasks, such as fire- vs. water-related objects and Western vs. Eastern objects. Visual deep neural networks (DNNs) such as ResNet-50 also showed poorer performance on these tasks, whereas humans ($n = 9$) performed almost perfectly on all tasks. The monkey performances for the three THINGS tasks in Fig. 2 were replotted using only the first 100 test stimuli to compare with the other tasks. See Supplementary Fig. 8a for the performance of other DNNs. Error bars indicate S.E.M. across images. (**c**) The DNN outputs separate concepts well in a low-dimensional space visualized via multi-dimensional scaling (MDS) for the tasks on which these networks performed well. For difficult tasks (e.g., Western vs. Eastern task), the outputs did not separate concepts. The plots show the results of ResNet-50. Other visual DNNs yielded similar patterns. (**d**) CLIP, a language-informed DNN, showed clearer clusters, even separating concepts that cannot be distinguished by the monkeys. (**e**) The strength of concept clustering measured as the geometric signal-to-noise ratio (SNR) (Sorscher et al. 2022) was consistently greater for DNNs than for lower-level models. The plot is the average of the 13 tasks that the monkeys could solve in a. See Supplementary Fig. 12 for the results of each task. The error bars indicate S.E.M. across tasks. All images were used either under CC0 license or Pixabay content license.*

using multi-dimensional scaling (MDS). For the tasks that the monkeys could solve, there was a clear boundary between the two concepts (Fig. 4c left and middle). By contrast, there was no clear boundary for the tasks that the monkeys could not solve (Fig. 4c right). Importantly, these visual DNNs were trained only to classify the ImageNet 1,000 categories, which do not contain many of the abstract concepts tested here. Thus, prominent abstract concepts appear to be present as high-level image features in a way that can be linearly extracted by either biological or artificial visual systems trained for general visual recognition. For comparison, we checked the output of CLIP (a language-informed model) and found that it produced more distinct clusters of responses even for concepts that could not be distinguished by the monkeys (Fig. 4d). An index of clustering (geometric signal-to-noise ratio; Sorscher et al. 2022; St-Yves et al. 2024) was significantly higher for DNNs than for low- and mid-level vision models (Fig. 4e; $p < 0.013$ for all comparisons, two-tailed *t*-test, Bonferroni corrected for multiple comparisons).

Discussion

Macaque monkeys serve as a primary model for investigating the human visual system, allowing characterization of neural mechanisms at the cellular level and providing insight into the evolutionary development of object vision (Orban et al. 2004; Yamins and DiCarlo 2016). Whereas their basic visual acuity and object discrimination performance are known to be similar to those of humans (Cherian et al. 2025; Koopman et al. 2017; Rajalingham et al. 2015), it has been understudied whether monkeys can classify visual objects according to abstract concepts such as animate, artificial, and big. We found that monkeys can be quickly trained to classify object images according to rules defined by such concepts. Their behavioral responses resembled human judgment (Fig. 3) and could not be explained by simple, low-level visual features (Fig. 2). But we do not intend to claim that monkeys internally possess the kinds of abstract concepts we operationally defined here. Instead, we suggest that the monkeys could solve the tasks because (1) there are high-level visual features diagnostic for solving these rules linearly embedded in the images (Figs. 2e, 4) and (2) primate brains have the capacity to extract these features and quickly associate them with actions. This capacity might be a building block for the formation of rich abstract concepts in the human mind.

We believe this is the first study that systematically investigated the ability of monkeys to classify a series of abstract object concepts using a binary categorization paradigm. It is well known that monkeys can classify animate vs. inanimate objects (Fabre-Thorpe 2003; Fabre-Thorpe et al. 1998; Roberts and Mazmanian 1988; Yetter et al. 2021) and that primate brains clearly encode animal appearance (Bracci et al. 2019; Kiani et al. 2007; Kramer et al. 2023; Kriegeskorte et al. 2008; Zhang et al. 2024). There are also many precedents that trained monkeys to classify specific object categories such as trees vs. non-trees (Vogels 1999) and cats vs dogs (Eldridge et al. 2018; Minamimoto et al. 2010). Extending these works, we tested more than 10 classification rules in one study with the same training schedule. This test battery revealed previously unknown classification abilities of monkeys, such as natural vs. artificial, mammal vs. non-mammal, big vs. small, and indoor vs. outdoor images, and identified the limits of their abilities as well (Fig. 4). These results

were further compared against large-scale human behavioral data and outputs of vision models.

It was a surprise to us that the monkeys learned each task rule quickly, although our tasks required them to discover a common rule for classification from many object images with different appearances, which marks a contrast with conventional visuomotor association tasks only using a few images (Cromer et al. 2011). Previous studies have also successfully trained monkeys on object categorization quickly (Eldridge et al. 2018; Minamimoto et al. 2010; Vogels 1999), suggesting that these classification skills are part of monkeys' natural behavioral repertoire. Our task paradigm, which requires direct image manipulation (Fig. 1a), could have also contributed to the fast learning by encouraging monkeys to pay attention to image features and providing them with more time for deliberate decision making, including the possibility of changes of mind (Supplementary Fig. 2; Kaufman et al. 2015). In addition to object recognition, our high-throughput behavioral testing with a binary decision-making paradigm can be useful in the future for assessing the visual skills of monkeys in a variety of tests from basic visual discrimination to complex image classification.

We also observed rather surprising consistency in the monkeys' learning curves across a variety of classification tasks. Our test battery included not only abstract rules but also some basic-level categorizations (monkeys vs. humans; Fig. 1g-i) and one vs. other tasks (Fig. 3i-k). In all the tasks, the monkeys monotonically improved their performance, reached a plateau at 2-4 days, and flexibly switched the rule in the subsequent sessions. We do not know what mechanisms underlie this learning pattern and flexibility; the flexibility might rely on prefrontal mechanisms, but the observed learning rate might align with the involvement of habit learning systems (Hikosaka et al. 2014). With respect to the similarity of learning speed across tasks, many studies posited that basic-level categories are most accessible due to the family resemblance of image features within a category (Grill-Spector and Kanwisher 2005; Rosch et al. 1976), whereas other studies have reported that superordinate categories are processed faster in the human brain (Clarke and Tyler 2015; Macé et al. 2009). But these differences in the speed might simply depend on the amount of evidence needed to be gleaned to make a decision (Mack and Palmeri 2011), and the underlying decision-making mechanisms might not qualitatively differ across different levels of object classification.

What do our findings tell us about the mechanisms of forming object concepts in the human mind? Since monkeys obviously do not benefit from human language and knowledge, their classification performance can be a testbed for the extent to which abstract concepts can be classified based on the visual processing capacity of primate brains. Importantly, visual DNNs, whose network weights were not explicitly trained using abstract concepts, were able to perform the same tasks using a linear classifier (Fig. 4; Jozwik et al. 2017; Kubilius et al. 2016; Peterson et al. 2018). Previous studies have also shown that other primates or even non-primate species, such as pigeons, can perform abstract visual classification (Lazareva et al. 2004; Pusch et al. 2023; Vonk and MacDonald 2002). Taken together, these observations suggest that some prominent object concepts are "embedded" as visual features in natural images in a way that can be readily extracted by both biological and artificial visual systems. This is consistent with a recent finding that DNNs trained only on children's visual experiences alone learn broad semantic categories (Orhan and Lake 2024). Thus, this embedded nature of object concepts may play an important role in the

formation of these concepts in humans during development (Spriet et al. 2022), although humans obviously acquire many more concepts that are difficult to classify based on image features (Fig. 4b).

By testing even more diverse stimuli and rules, our task paradigm can contribute to a better understanding of what computations occur in the ventral visual pathway (Schrimpf et al. 2018) and how its output is read out by downstream areas (Majaj et al. 2015). A comparison of the monkey responses and DNN outputs revealed that networks with different internal architectures performed similarly well (Fig. 2e), echoing a recent finding that DNN architecture is not a primary factor in predicting ventral-pathway responses to natural images (Conwell et al. 2022). To understand what features monkeys rely on to perform various concept classification tasks, it is important to use more controlled image sets specifically designed to dissociate different visual recognition algorithms (Bowers et al. 2023). This might reveal fundamental differences in classification strategies between humans and monkeys, which we did not identify using only natural images and a small number of control images (Fig. 3).

How humans acquire and recognize object concepts has long been a central question in various fields of research. Furthermore, recent advances in deep neural network models have sparked debate about whether these models have acquired internal representations relevant to the inner workings of the human brain. We expect that studying the behavioral and neural mechanisms of non-human primates with the kinds of task paradigms we have developed can provide unique insights into how object concepts can be formed in the human mind.

Acknowledgements

We thank Takafumi Minamimoto and Cheng Xue for the comments on earlier versions of the manuscript. We thank Roozbeh Kiani for providing natural object images and also thank the Japan Monkey Centre and the Center for the Evolutionary Origins of Human Behavior, Kyoto University, for providing monkey images used in the experiments. This work was supported by Strategic Priority Research Program of the Chinese Academy of Sciences (XDB1010202), the National Science and Technology Innovation 2030 Major Program (Grant No. 2021ZD0203703), National Natural Science Foundation of China (Grant No. 32371077 and No. W2432019), and the National Natural Science Fund for Excellent Young Scientists Fund Program (overseas).

Competing interests

The authors declare no competing interests.

Data and code availability

Data and code used in this study will be made publicly available upon official publication.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Bracci, S., Ritchie, J. B., Kalfas, I., & de Beeck, H. P. O. (2019). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, *39*(33), 6513–6525.

Butler, J. L., & Kennerley, S. W. (2019). Mymou: A low-cost, wireless touchscreen system for automated training of nonhuman primates. *Behavior Research Methods*, *51*(6), 2559–2572.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142.

Cherian, T., Jacob, G., & Arun, S. (2025). Do monkeys see the way we do? qualitative similarities and differences between monkey and human perception. *bioRxiv*, 2025–01. https://doi.org/10.1101/2025.01.30.635614

Clarke, A., & Tyler, L. K. (2015). Understanding what we see: How we derive meaning from vision. *Trends in Cognitive Sciences*, *19*(11), 677–687.

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, *8*(11), 2179–2193.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*. https://doi.org/10.1101/2022.03.28.485868

Cromer, J. A., Machon, M., & Miller, E. K. (2011). Rapid association learning in the primate prefrontal cortex in the absence of behavioral reversals. *Journal of Cognitive Neuroscience*, *23*(7), 1823–1828.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is

worth 16x16 words: Transformers for image recognition at scale. *arXiv*. https://doi.org/10.48550/arxiv.2010.11929

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816.

Eldridge, M. A., Matsumoto, N., Wittig, J. H., Masseau, E. C., Saunders, R. C., & Richmond, B. J. (2018). Perceptual processing in the ventral visual stream requires area TE but not rhinal cortex. *eLife*, *7*, e36310.

Fabre-Thorpe, M. (2003). Visual categorization: Accessing abstraction in non–human primates. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1435), 1215–1223.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, *9*(2), 303–308.

Fize, D., Cauchoix, M., & Fabre-Thorpe, M. (2011). Humans and monkeys share visual representations. *Proceedings of the National Academy of Sciences, USA*, *108*(18), 7635–7640.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

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

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv*. https://doi.org/10.48550/arxiv.1512.03385

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS One*, *14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Heidari-Gorji, H., Ebrahimpour, R., & Zabbah, S. (2021). A temporal hierarchical feedforward model explains both the time and the accuracy of object recognition. *Scientific Reports*, *11*(1), 5640.

Hesse, J. K., & Tsao, D. Y. (2020). The macaque face patch system: A turtle's underbelly for the brain. *Nature Reviews Neuroscience*, *21*(12), 695–716.

Hikosaka, O., Kim, H. F., Yasuda, M., & Yamamoto, S. (2014). Basal ganglia circuits for reward value–guided behavior. *Annual Review of Neuroscience*, *37*(1), 289–306.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Kar, K., & DiCarlo, J. J. (2024). The quest for an integrated set of neural mechanisms underlying object recognition in primates. *Annual Review of Vision Science*, *10*, 91–121.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2015). Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex. *eLife*, *4*, e04677.

Kawaguchi, K., Pourriahi, P., Seillier, L., Clery, S., & Nienborg, H. (2019). Easily adaptable head-free training system of macaques for tasks requiring precise measurements of eye position. *bioRxiv*. https://doi.org/10.1101/588566

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309.

Kim, T., Bair, W., & Pasupathy, A. (2019). Neural coding for shape and texture in macaque area V4. *Journal of Neuroscience*, *39*(24), 4760–4774.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, *74*(6), 1114–1124.

Koopman, S. E., Mahon, B. Z., & Cantlon, J. F. (2017). Evolutionary constraints on human object perception. *Cognitive Science*, *41*(8), 2126–2148.

Kramer, L. E., Chen, Y.-C., Long, B., Konkle, T., & Cohen, M. R. (2023). Contributions of early and mid-level visual cortex to high-level object categorization. *bioRxiv*. https://doi.org/10.1101/2023.05.31.541514

Kriegeskorte, N., & Mur, M. (2012). Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Lazareva, O. F., Freiburger, K. L., & Wasserman, E. A. (2004). Pigeons concurrently categorize photographs at both basic and superordinate levels. *Psychonomic Bulletin & Review*, *11*(6), 1111–1117.

Liu, N., Kriegeskorte, N., Mur, M., Hadj-Bouziane, F., Luh, W.-M., Tootell, R. B., & Ungerleider, L. G. (2013). Intrinsic structure of visual exemplar and category representations in macaque brain. *Journal of Neuroscience*, *33*(28), 11346–11360.

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95.

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision*, *17*(6), 20–20.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, USA*, *115*(38), E9015–E9024.

Luo, T., Xu, M., Zheng, Z., & Okazawa, G. (2025). Limitation of switching sensory information flow in flexible perceptual decision making. *Nature Communications*, *16*(1), 172.

Macé, M. J.-M., Joubert, O. R., Nespoulous, J.-L., & Fabre-Thorpe, M. (2009). The time-course of visual categorizations: You spot the animal faster than the bird. *PloS one*, *4*(6), e5927.

Mack, M. L., & Palmeri, T. J. (2011). The timing of visual object categorization. *Frontiers in Psychology*, *2*, 165.

Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology*, *60*(1), 27–51.

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in Cognitive Sciences*, *15*(3), 97–103.

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

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*(1), 25–45.

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

Murphy, G. (2004). *The big book of concepts*. MIT press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289.

Okazawa, G., & Kiani, R. (2023). Neural mechanisms that make perceptual decisions flexible. *Annual Review of Physiology*, *85*(1), 191–215.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque v4. *Proceedings of the National Academy of Sciences, USA*, *112*(4), E351–E360.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, *8*(7), 315–324.

Orhan, A. E., & Lake, B. M. (2024). Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, *6*(3), 271–283.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), e27.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*, 49–70.

Pusch, R., Clark, W., Rose, J., & Güntürkün, O. (2023). Visual categories and concepts in the avian brain. *Animal Cognition*, *26*(1), 153–173.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*. https://doi.org/10.48550/arxiv.2103.00020

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, *35*(35), 12127–12136.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55.

Ramezanpour, H., Giverin, C., & Kar, K. (2024). Low-cost, portable, easy-to-use kiosks to facilitate home-cage testing of nonhuman primates during vision-based behavioral tasks. *Journal of Neurophysiology*, *132*(3), 666–677.

Roberts, W. A., & Mazmanian, D. S. (1988). Concept learning at different levels of abstraction by pigeons, monkeys, and people. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(3), 247.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Santos, L. R., Hauser, M. D., & Spelke, E. S. (2001). Recognition and categorization of biologically significant objects by rhesus monkeys (macaca mulatta): The domain of food. *Cognition*, *82*(2), 127–155.

Schrier, A. M., & Brady, P. M. (1987). Categorization of natural stimuli by monkeys (macaca mulatta): Effects of stimulus set size and modification of exemplars. *Journal of Experimental Psychology: Animal Behavior Processes*, *13*(2), 136.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. https://doi.org/10.1101/407007

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*. https://doi.org/10.48550/arxiv.1409.1556

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences, USA*, *119*(43), e2200800119.

Spriet, C., Abassi, E., Hochmann, J.-R., & Papeo, L. (2022). Visual object categorization in infancy. *Proceedings of the National Academy of Sciences, USA*, *119*(8), e2105866119.

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between integration and non-integration strategies in perceptual decision making. *eLife*, *9*, e55365.

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2024). THINGSplus: New norms and metadata for the things database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, *56*(3), 1583–1603.

St-Yves, G., Kay, K., & Naselaris, T. (2024). Variation in the geometry of concept manifolds across human visual cortex. *bioRxiv*. https://doi.org/10.1101/2024.11.26.625280

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. part 1: Behavioural study. *European Journal of Neuroscience*, *11*(4), 1223–1238.

Vogels, R. (2022). More than the face: Representations of bodies in the inferior temporal cortex. *Annual Review of Vision Science*, *8*(1), 383–405.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511.

Vonk, J., & MacDonald, S. E. (2002). Natural concepts in a juvenile gorilla (gorilla gorilla gorilla) at three levels of abstraction. *Journal of the Experimental Analysis of Behavior*, *78*(3), 315–332.

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

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA*, *111*(23), 8619–8624.

Yetter, M., Robert, S., Mammarella, G., Richmond, B., Eldridge, M. A., Ungerleider, L. G., & Yue, X. (2021). Curvilinear features are important for animate/inanimate categorization in macaques. *Journal of Vision*, *21*(4), 3–3.

Yoshikubo, S. (1985). Species discrimination and concept formation by rhesus monkeys (macaca mulatta). *Primates*, *26*, 285–299.

Zhang, Z., Hartmann, T. S., Born, R. T., Livingstone, M. S., & Ponce, C. R. (2024). Brain feature maps reveal progressive animal-feature representations in the ventral stream. *bioRxiv*. https://doi.org/10.1101/2024.11.24.625066

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.

Methods

Subjects and apparatus . . . 24
Monkey experiments . . . 24
Human experiments . . . 25
Object drag task . . . 25
Image sets . . . 26
Shape images for pretraining . . . 26
Background-cropped image sets (Figure 1) . . . 26
Large-scale THINGS image sets (Figure 2) . . . 27
Modified control image sets (Figures 2d, 3g-h) . . . 28
Image sets for more conceptual tasks (Figure 4) . . . 29
Outdoor vs. indoor scenes . . . 29
Big vs. small objects . . . 29
Fire- vs. water-related objects . . . 29
Eastern- vs. Western-culture objects . . . 29
Texform images (Supplementary figure 5) . . . 30
Initial monkey training procedure . . . 30
Monkey experimental procedures . . . 31
Object drag tasks using grayscale image sets (Figure 1) . . . 31
Object drag tasks using large-scale THINGS image sets (Figure 2) . . . 31
Tasks using modified control images (Figures 2d, 3g-h) . . . 31
One-versus-other tasks (Figure 3i) . . . 32
More abstract classification tasks (Figure 4) . . . 32
Random stimulus-response association tasks (Supplementary figure 3) . . . 32
Human experimental procedures . . . 33
Object drag tasks using grayscale image sets (Figure 3) . . . 33
Object drag tasks with generalization images (Figure 4) . . . 33
Multi-object arrangement task (Figure 3l) . . . 34
Data analysis . . . 34
Choice accuracy and reaction time . . . 34
Construction of dissimilarity matrix and dendrogram from one-versus-other tasks (Figure 3k-l) . . . 35
Analysis of multi-object arrangement task for humans (Figure 3l) . . . 35
Model comparison . . . 36
Category classification and behavioral fitting procedures (Figure 2e) . . . 36
Model details . . . 37
V1-like model . . . 37
Luminance+color model . . . 37
Gist+SF model . . . 37
Texture model . . . 38
Contour model . . . 38
Deep neural network (DNN) models . . . 38
Calculation of the geometric signal-to-noise ratio (SNR) . . . 38

Comparison of monkey and human behavior (Figure 3) 39
Metric based on the drift-diffusion model (DDM) 39
Inverse efficiency metric 40

Subjects and apparatus

Monkey experiments

We collected behavioral data from three adult rhesus monkeys (*Macaca mulatta*; monkey Ju, female, age 13; monkey El, female, age 12; monkey Ol, male, age 11) in various object classification tasks. These monkeys were raised in a typical laboratory environment and had no prior experience performing relevant behavioral tasks. The dataset collected in this study consisted of 26 tasks including subtasks, amounting to $\sim$315K trials from the three monkeys. Experimental procedures conformed to the National Institutes of Health *Guide for the Care and Use of Laboratory Animals* and were approved by the Animal Care Committee at Center for Excellence in Brain Science and Intelligence Technology (Institute of Neuroscience), Chinese Academy of Sciences.

Experiments were conducted using a custom-built touchscreen device attached to each monkey's cage (Supplementary Fig. 1a), similar to those developed by other laboratories (Butler and Kennerley 2019; Ramezanpour et al. 2024). The device box consisted of a commercial touchscreen (AOSIMAN, ASM-125UCT, 12.5 inches, size 287 mm x 186 mm; 60 Hz refresh rate; 3840 $\times$ 2160 pixel screen resolution), a reward juice delivery system with a TTL-triggered water pump, a monitoring camera, and a control computer (MOREFINE S500+AMD R9-5900HX). The monkey reached the touchscreen through a hole in the front panel attached to the cage. After a few training sessions, the monkeys consistently used the same arm to reach the screen (monkey Ju: left arm, monkeys El and Ol: right arm). The monkey viewed the screen through another hole, which opened at approximately the level of the monkey's head. Below the hole was a reward juice outlet in a "cone"-shaped socket of a size suitable to cover the monkey's mouth (Kawaguchi et al. 2019). The monkeys performed the tasks while their mouth was in this socket and thus had a stable head position, which was confirmed by monitoring them online with a camera. Under these conditions, the viewing distance to the monitor was approximately 15-18 cm. The task was controlled by a custom-made MATLAB (MathWorks, MA, USA) program with Psychtoolbox.

A daily behavioral session typically lasted 1-2 hours in the monkey's home cage, and the touchscreen device was attached to the cage during this period. We found that limiting time can better control monkey's engagement and behavioral performance. During a session, the monkeys usually continuously performed the tasks without interruption, and we remotely adjusted the task parameters if needed.

Human experiments

We recruited in total 33 human participants (20-40 years old; 10 males and 23 females, students or employees at the Chinese Academy of Sciences). Our participant sampling did not consider sex, as it was considered unlikely to influence the results of this study. They were assigned to perform different tasks, as explained in the "Human experimental procedures" section below. All participants had normal or corrected-to-normal vision and were naïve to the purpose of the experiments. Written informed consent was obtained from the participants. All experimental procedures were approved by the Institutional Review Board of the Center for Excellence in Brain Science and Intelligence Technology (Institute of Neuroscience), Chinese Academy of Sciences.

During the experiments, the participants were seated in a height-adjustable chair in a half-lit room, with their chin and forehead supported by a chinrest. The chinrest ensured a stable viewing distance (37 cm) to a touchscreen monitor (LENOVO, T24t-20, 23.8 inches; 60 Hz refresh rate; 1920 x 1080 pixel screen resolution). Although the view distances and touchscreen were different from those used in monkey experiments to accommodate the differences in their arm lengths, we maintained similar visual angles of the object images (see below). The other task parameters were also kept identical to those of the monkey experiments to ensure a direct comparison.

Object drag task

We developed a binary object categorization task ("object drag task") for the touchscreen device (Fig. 1a). During the task, the participants viewed an object stimulus and moved it to one of two target boxes by swiping it with their finger. The two target boxes were associated with distinct object concepts (e.g., animate vs. inanimate objects). We considered that touching and moving an object to a target would provide a sense of direct interaction with the object and facilitate the learning of the association.

Each trial started when the participants touched a red fixation dot. Following this, a stimulus ($10 - 11$ deg) appeared together with two gray square target boxes. The positions of the stimulus and target boxes were adjusted for each participant so that they could move a stimulus smoothly to both targets with similar amounts of time. In the monkey experiments, the distance between the stimulus and target boxes was $18 - 24$ deg, and the distance between the two target boxes was $24 - 37$ deg. Following the stimulus presentation, the participants were required to touch the stimulus within 1 s. After the touch, the participants were required to drag the stimulus to one of the two target boxes within 5 s. In the case of timeout, the trial was aborted (1.3 % of trials for monkeys). When the image reached one of the two boxes, the image and the target boxes disappeared, and instead, the same image appeared at the correct target location, providing visual feedback to the participants. At the same time, an auditory tone indicated correct and incorrect responses. The monkeys received a juice reward for a correct response and a short timeout (1-2 s) for an incorrect response.

A rule for a given task (e.g., animate vs. inanimate) did not change within a daily experimental

session for the monkeys (see below for rules). The same rule usually lasted for one to two weeks until the participants learned the task and completed a generalization test (Fig. 1b). At the start of a new rule, they were not informed of the rule change, but the new task involved different stimuli, which may have reduced the interference from the previous task and facilitated the learning of the new rule.

Image sets

Shape images for pretraining

To train the monkeys on the structure of the task prior to the main data collection, we used nine groups of shape images created using a custom MATLAB program (Supplementary Fig. 1b). The program first generated a three-dimensional (3D) model of a sphere and then randomly added multiple Gaussian protrusions of different heights and widths to the sphere to generate amorphous shapes. By starting with different randomization seeds, we generated many basic shapes ("basic" set (1)). We added various features to these basic shapes to create different groups as follows: (2-3) "Red" and "Green" sets consisting of shape stimuli colored in red and green. (4) "Sharp" set created by narrowing the Gaussian width. (5) "Square" set created by transforming the sphere into a cube before adding the Gaussian protrusions. (6-7) "Texture 1" and "Texture 2" sets created by drawing bark and brick-wall textures on the surface, respectively. (8-9) "Horizontal" and "Vertical" sets created by stretching the shapes either horizontally or vertically. There were 50 random shape images in each category.

Background-cropped image sets (Figure 1)

In the first series of our experiments (Fig. 1), we used grayscale object images with the background removed (examples shown in Fig. 1b, g). Some of the images (5%) were obtained from the existing databases (Kiani et al. 2007), whereas the remaining images were obtained by manually removing background from natural photographs of objects obtained through search engines such as Bing, Pixabay, and Google Image. We also obtained photographs of monkeys from the Japan Monkey Centre and the Center for the Evolutionary Origins of Human Behavior, Kyoto University, which were also processed by removing the background. Images were converted to grayscale because the distribution of colors was highly skewed in some objects (e.g., monkey images are always brownish), and this could present a strong low-level cue for classification. Because some original images do not allow redistribution, we replaced them with similar images with the Creative Commons Zero (CC0) license or other royalty-free licenses when showing in the figures. The sources of the human images shown in Figs. 1, 3 and Supplementary Fig. 4 are PurePNG (https://purepng.com/; authors: LPuo, hadiaaltaf, Momo).

When collecting images for these sets, we focused on eight object groups: humans, monkeys, non-primate mammals, non-mammal (reptiles/amphibians), food, plants, tools/everyday objects,

and electronics. Although these eight groups span across different levels of category hierarchy (e.g., "monkey" can be considered a basic category, whereas "food" is a superordinate category), we considered that these groups covered a range of object categories ideal for behavioral experiments with monkeys. First, these groups address biological significance and familiarity to experimental monkeys. For example, many images from the "human", "monkey", and "food" groups are familiar and relevant to monkeys, whereas they should have limited experience with "reptile/amphibian" and "electronics". Second, these eight can be divided into four animate and inanimate categories. The four inanimate groups can be further divided into natural (food and plants) and artificial (tools/everyday objects and electronics) objects. The four animate groups can be divided into primate (human and monkey) and non-primate (non-primate mammal, reptile/amphibian) species. Thus, we considered that these eight groups provided a well-balanced structure specifically for the monkey experiments.

Using these eight object groups, we created six tasks (Fig. 1b, g): animate vs. inanimate, natural vs. artificial, non-primate mammal vs. non-mammal (reptile/amphibian), human vs. monkey, monkey vs. non-primate mammal, tool/everyday object vs electronics. In each task, 60-96 images were used to train the monkeys, and 20-96 images were used to test their performance on novel stimuli after learning (generalization test) (Fig. 1b). Training images could have a small number of overlaps across tasks, whereas the generalization images were never shown to the monkeys in any previous tasks. In each training set, we attempted to sample a wide range of objects, but at the same time, often included two images from the same basic-level categories as we sought to compare their performance early on.

Large-scale THINGS image sets (Figure 2)

To complement the experiments using the controlled but small stimulus sets above, we created large-scale image sets based on the THINGS database developed by Hebart et al. (Hebart et al. 2019, 2020; Stoinski et al. 2024). The THINGS database consists of 1,814 concrete object concepts (such as dog, spider, lemon, and table; referred to as "object categories" in this paper to avoid confusion), each of which contains at least 12 natural photographs. The selection of object categories and associated images are fully human-curated, achieving comprehensive coverage of common object categories with high-quality images. The images have color and include backgrounds that usually contain no other distinct objects. There were still a fraction of images that contained multiple objects (such as an image of a tool with a hand), and we manually removed these images to avoid causing confusion for the monkeys.

For the purpose of our experiments, we selected a subset of object categories from the THINGS database and created three tasks (Fig. 2a): animate vs. inanimate (373 object categories, 3,683 images in total), natural vs. artificial objects (230 object categories, 1,915 images in total), and mammal vs non-mammal animals (143 object categories, 1,668 images in total). These object categories are much more diverse than the ones we had in the grayscale image sets, encompassing many objects that experimental monkeys were unlikely to have experience with. Among the images, we selected 100 images (50 images per target; no overlap in object categories) for training the monkeys on each task. During the generalization test, we showed the remaining images only

once each. Since there were multiple images in each object concept, the generalization images included those from the same object categories as one of the training images ("old category" in Fig. 2a) or from different object categories ("new category"). Due to copyright issue, the images shown in the figures are replaced with similar images with the CC0 license from the THINGSplus database (Stoinski et al. 2024) or images with the Pixabay Content license from Pixabay.

To inspect the patterns of monkey behavior, we further classified these objects into high-level object categories such as "bird", "plant", and "fruit" (Fig. 2c). These high-level categories were mostly taken from the THINGS database (Hebart et al. 2019), which uses human curation to determine them. In addition, we included several high-level categories that we defined ourselves (mammals, non-mammals, natural, artificial, reptiles/amphibians, fish, and mollusks/crustaceans) as they could be informative for assessing monkey performance.

Modified control image sets (Figures 2d, 3g-h)

After performing the animate vs. inanimate task using the grayscale, background-cropped images (Fig. 1), we tested whether the monkeys could classify images outside the range of natural photographs. We created three control image sets. First, an "outline" image set was created by extracting the outlines of the object images using *bwboundaries* function in MATLAB. The contour was colored black, while its interior was filled with the background color of the task screen (gray). Second, we created a "silhouette" image set, where the inside of a contour was also filled with black. Third, we created a "cartoon" image set by selecting cartoon pictures of object images from a publicly available database ("Multipic" database; Duñabeitia et al. 2018).

In all the sets, we selected images whose object categories partially overlapped with those in the grayscale training image set, but we never used the same images. For example, "outline" or "silhouette" images were created from object images that were not shown during the training to ensure that the monkeys could not solve the task by remembering the outlines of specific object images. Each set contained 60 images.

We also presented two control image sets after performing the categorization of large-scale THINGS images. First, during the natural vs. artificial task, we showed THINGS images converted into grayscale to confirm that the monkeys did not simply rely on color. Second, during the mammal vs. non-mammal task, we presented THINGS images whose texture was distorted by using *crystallization* function (cell size, 10) in Photoshop 2024 (Adobe Inc., San Jose, CA, USA). We had 400 images for each set, selected from the generalization image set. These images were shown once during the generalization test, but we consider it is unlikely that the monkeys remembered specific images to solve these control tasks because they were shown only once and mixed with many other images.

Image sets for more conceptual tasks (Figure 4)

To further challenge the monkeys' ability to learn abstract rules, we created additional conceptual tasks (Fig. 4). These tasks included indoor vs. outdoor scenes, two versions of big vs. small objects, fire- vs. water-related objects, and Western- vs. Eastern-culture objects. Each task used 90-120 training images and 90-120 generalization images.

Outdoor vs. indoor scenes We selected scene images from Zhou et al. (2017). The indoor scenes included pictures of a variety of buildings, including small scenes such as bedrooms and kitchens and also spacious scenes such as stadiums and concert halls. The outdoor scenes included both natural scenes, such as forests, and urban scenes, such as buildings. Because the sky and natural objects (e.g., a tree) could be strong cues indicating outdoors, we preferentially selected images that did not contain them (e.g., the front face of a building). Nevertheless, outdoor scenes tend to have certain colors (greenish). Therefore, we created a grayscale version of the generalization images as a control and confirmed that the monkeys also performed well without color (73 % correct).

Big vs. small objects In this task, the monkeys had to classify objects based on their physical size. Objects that are typically bigger or smaller than the size of a chair were classified as big or small, respectively (Konkle and Oliva 2012; Long et al. 2016). We created two sets of stimuli. The first set used images from the THINGS database, which had a natural background. Small objects included many household items such as spoons and nails, and big objects included furniture (such as bed and ovens), vehicles, and others. We ensured that pictures of most big objects were taken indoors so that this categorization could not be confused with outdoor vs. indoor. The second set used images from the big vs. small images created by Konkle et al. (2012), which had no background. A comparison of these two sets allowed us to confirm that the monkey performance did not depend on a specific choice of images.

Fire- vs. water-related objects To create abstract rules that are unlikely to be correlated with any visual features, we selected object images that are semantically associated with fire and water from the THINGS database. The fire-related category included objects such as stoves, cigarettes, coal, fire trucks, and radiators. The water-related category included objects such as hoses, bathtubs, paddles, and squirt guns. The images of fire or water itself were avoided as they have distinctive features.

Eastern- vs. Western-culture objects As another abstract categorization task, we created an image set of objects associated with Eastern and Western cultures. We collected the images mainly by using Internet search engines, because a sufficient number of images could not be found in the THINGS database. The Eastern category included objects such as Chinese lanterns, dumplings,

old pavilions, bamboo, and chopsticks. The Western category included objects such as crowns, donuts, the Eiffel Tower, tulips, and violins.

Texform images (Supplementary figure 5)

Texforms are synthetic images generated by matching the mid-level texture features of natural object images (Long et al. 2016, 2017, 2018). These features consist of the magnitudes and correlation structures of V1-like oriented bandpass filter outputs and some other image statistics at each local image region (Freeman et al. 2013; Portilla and Simoncelli 2000). Humans can reliably classify the animacy of original objects from texform images (Long et al. 2017); thus we tested whether monkeys could generalize the animate vs. inanimate task to texform images. We used the original object and texform images from Long et al. (2018).

We first created a set of original object images (30 each for animate and inanimate images) to confirm that the monkeys could successfully classify these original images (Supplementary Fig. 5a). We then created a texform set (30 each for animate and inanimate; Supplementary Fig. 5b). The texform images were chosen such that they were not made from the original images we have shown because the monkeys might have remembered the original images. After testing the generalization performance of this texform set, we trained the monkeys on this texform classification for 7-8 sessions (Supplementary Fig. 5c). Finally, we created another texform set (30 each for animate and inanimate; Supplementary Fig. 5d) and tested whether the monkeys could generalize the learned texform classification rule to new texform images. The set contained both big and small objects for animate and inanimate images, and we evenly distributed them for each set.

Initial monkey training procedure

Prior to the main data collection, the monkeys were trained on the task structure using shape images (see the "Shape images for pretraining" section above). The shape images consisted of night categories, differing in color, shape, and texture (Supplementary Fig. 1b) and were irrelevant to the abstract concepts tested in our main experiments. We first trained monkeys to touch and move a shape image to one target box on the screen ($<$ 1 week). We then trained "basic" vs. "red" category classification using a match-to-sample design. In this design, we presented two target images (chosen from "basic" and "red"; they were always placed on the same two sides) and trained the monkeys to move a stimulus (either "basic" or "red") to the same target image. During training, we adjusted the opacity of the wrong target image from being transparent (invisible) to full contrast to gradually increase task difficulty. After the monkeys learned the match-to-sample task, we made a transition to our main task by gradually superimposing gray covers on two target images. After the covers became full contrast, the task became equivalent to our main design (Fig. 1a). This entire training of "basic" vs. "red" classification took 9-11 days (Supplementary Fig. 1c). We then trained monkeys on a series of shape categorization tasks (e.g., "green" vs. "sharp", "square" vs. "texture 1"; Supplementary Fig. 1b-c) to ensure that they could learn novel stimulus-target associations in our main design. The monkeys learned these tasks rapidly within a few days

(Supplementary Fig. 1c).

Monkey experimental procedures

Object drag tasks using grayscale image sets (Figure 1)

We performed six tasks using grayscale image sets (Fig. 1b, g). The procedure for each task consisted of training and generalization test sessions. During the training sessions, an image from the training set was pseudo-randomly presented to the monkey in each trial. Training lasted for 4-6 sessions (days) depending on the monkey's learning speed. In the early training phase of each task, we had a portion of trials in which we showed only the correct target box until the monkey performance reached above $\sim 70\%$. These trials were used to maintain the monkey's engagement and were excluded from the data analysis. When monkeys showed a strong choice bias for one target, we penalized the bias by showing a stimulus associated with the opposite target after three consecutive errors on that side. However, these penalty trials were rare (0.4 % of trials) and were also excluded from the data analysis. Immediately after the training sessions, we had a generalization session, in which we started to show generalization images in 50% of the trials. We analyzed only the trials in which these generalization images were presented for the first time.

Object drag tasks using large-scale THINGS image sets (Figure 2)

We performed three tasks using large-scale THINGS image sets. For each task, we first confirmed that the monkeys perform well on the same task with the grayscale image set. We then started to introduce 100 training images from the THINGS set (Fig. 2a). After we confirmed that the monkey performance was stable for the 100 training images, we gradually introduced generalization images. Each generalization image was shown only once. We controlled the proportion of trials in which generalization images were presented, gradually increasing it up to 50%. In the other 50% of the trials, the training images were repeatedly shown to ensure that the monkeys continued to follow the same rule. We continued the sessions until all the generalization images were shown, which took $7-12$ sessions for the animate vs. inanimate task and $4-6$ sessions for the other tasks.

Tasks using modified control images (Figures 2d, 3g-h)

We presented images modified from natural photographs, such as object silhouettes in control sessions. See the "Modified control image sets" section above for the generation of these images. After the monkeys learned the animate vs. inanimate task using the grayscale image set (Fig. 1b-d), we presented silhouette images, outline images, and cartoon images separately in different sessions. In each session, we gradually mixed these control images with the training images until the proportion of the control set reached 50%. Similarly, for the natural vs. artificial and mammal vs. non-mammal tasks using the THINGS database (Fig. 2a), we presented grayscale images and

texture-modified images (Fig. 2d) in a separate session. We analyzed only the trials in which each of the control images was presented for the first time.

One-versus-other tasks (Figure 3i)

To infer the hierarchical structure of category confusion in monkey behavior, we performed one-versus-other tasks with one monkey (monkey Ju). We created an image set consisting of eight object groups (humans, monkeys, non-primate mammal, non-mammal (reptiles/amphibians), food, plants, tools/everyday objects, and electronics; 50 grayscale, background-cropped images per category; 400 images in total) and trained the monkey to perform a one-versus-other task for each category, in which one target was associated with that category and the other target was associated with all others (Fig. 3i). Thus, the monkey was trained on a total of eight tasks. During each task, images from the target category and the other categories appeared with equal probability, although there were seven times more images in the other categories. Training on each task took 4-6 days. We analyzed the data from the last 3-4 sessions to construct a confusion matrix (Fig. 3k).

More abstract classification tasks (Figure 4)

Experiments on more classification tasks (Fig. 4; outdoor vs. indoor scenes, big vs. small objects, fire- vs. water-related objects, and Eastern- vs. Western-culture objects) also largely followed the procedure for the other tasks described above (Fig. 1b). Training for each task continued until the monkey performance plateaued but was terminated after six days regardless of their final performance to fairly compare their performance across tasks with similar training durations. The average performances for the outdoor vs. indoor, fire- vs. water-related, and Western vs. Eastern-culture tasks at the last day of training was 91%, 77%, and 88% correct, respectively. The big vs. small task had two versions: one using the THINGS database and one using the images from Konkle et al. (2012). We started with the THINGS set (final performance during training: 94%) and, after its generalization test, switched to training on the Konkle set (final performance during training: 93%).

Random stimulus-response association tasks (Supplementary figure 3)

We conducted two versions of control experiments in which monkeys were trained to associate random object images with targets (Supplementary Fig. 3) to check how the absence of a common conceptual rule impaired their performance. In one version (Supplementary Fig. 3a), the associations of images and targets were randomized such that even images of the same object category could be associated with different targets. In the second version (Supplementary Fig. 3b), two targets were associated with random concrete object concepts (e.g., “deer”, “banana”; we refer to them as “concrete categories” here to avoid confusion), but images in each concrete category were associated with the same target. Thus, if monkeys remembered the association of concrete

categories and targets, they could solve the task and generalize the rule to novel stimuli selected from the same concrete categories.

Both versions of the task used the same number of grayscale object images as the animate vs. inanimate task (96 images in the training set; Fig. 1b) and followed the same training procedure. The training image set for the second version (Supplementary Fig. 3b) contained 16 concrete categories for each target (2-6 images per concrete category, 48 images per target). As in the other grayscale object tasks (Fig. 1), these concrete categories were sampled from humans, monkeys, non-primate mammals, reptiles/amphibians, food, plants, tools/everyday objects, and electronics. We also constructed a generalization image set with the same number of images and object categories. All the images used in these control experiments differed from those used in the main tasks.

Human experimental procedures

Object drag tasks using grayscale image sets (Figure 3)

To compare the behavioral performance of humans and monkeys for the grayscale, background-cropped image sets, we performed six grayscale image tasks (Fig. 1b, g) with human participants. Seven participants were assigned to each task. In each trial, an image from the training image set was randomly presented, and the participants had to choose a target as in the monkey experiments. The participants were not informed of the categorization rule and had to infer it from the feedback, but they quickly understood the rule in all tasks (Fig. 3a). Then, each participant continued the task for $\sim$1,500 trials, resulting in 15-25 repetitions per image. This number of repetitions allowed us to fit the drift-diffusion model, in which category sensitivity was estimated for each image based on reaction time distributions (Fig. 3c-d). We also conducted the animate vs. inanimate task using three control image sets (cartoons, silhouettes, outlines; Fig. 3g, h). Four participants were recruited for each set, and each participant performed $\sim$1,500 trials. In total, 83,416 trials were collected across all tasks and participants.

Object drag tasks with generalization images (Figure 4)

We also measured the generalization performance of the human participants to novel stimuli in the various tasks we tested with the monkeys (15 tasks in total; Fig. 4). Nine participants were assigned to each task. For each task, we first asked the participants to categorize the training image set. We continued training until the participants reached 90% accuracy (20 trial sliding window) and viewed all the training images (65-183 trials). We then showed the generalization images only once. The same training and generalization images were used as those in the monkey experiments. In total, 13,754 trials for the generalization images were collected across all tasks and participants.

Multi-object arrangement task (Figure 3l)

To estimate the hierarchical structure of category similarity for humans, we performed a multi-object arrangement task (Kriegeskorte and Mur 2012) with six participants. In this task, multiple object images were displayed simultaneously on a screen, and the participants sorted them in a 2D arena such that their perceived similarity of each pair of images could be reflected by the distance of the pair in the arena. From the image locations they chose, we can generate a dissimilarity matrix for images (Kriegeskorte and Mur 2012). We used this task instead of the one-versus-other task used to assess monkey behavior because the one-versus-other task requires many trials to estimate category confusion in humans as it had to rely on reaction time distributions (Fig. 3c). In addition, the multi-object arrangement task requires human participants to directly report their perceived object similarity, allowing us to compare this direct report with monkey behavior. In total, 812 trials were collected across the six participants.

The task sequence largely followed that of Kriegeskorte et al. (2012). During each trial, after a participant held a fixation dot with a finger, a white circular arena (29 deg) appeared in the center of the screen. At the same time, 8-24 object images were shown at random positions surrounding this arena. The participants needed to move all the images within the arena and position them so that similar objects were placed nearby. After sorting, they pressed a button to indicate the end of the trial. We used the same set of images as in the monkeys' one-versus-other task, which had 8 categories and 50 images each. In the first 10 trials of each session, 24 images (8 categories, 3 images per category) were randomly selected, and participants had to arrange them within 100 s. In the following 70 trials, 8 images (4 categories, 2 images per category) were randomly selected (within 60 s). In these 70 trials, every possible combination of four categories were presented without bias. Showing these subsets of categories is considered a better approach to quantify details of the pairwise similarity, although we did not use the algorithm to optimally choose the subsets as Kriegeskorte et al. (2012) did.

Data analysis

Choice accuracy and reaction time

When calculating the choice accuracy of a session (Figs. 1d, 1h, 3j, Supplementary Figs. 3c, 5c), we used all trials after removing a subset of trials early in the training sessions, in which we showed only the correct target to maintain the monkeys' motivation (see "Monkey experimental procedures" above). These one-target trials were excluded from all analyses except for the inset in Fig. 1d, in which we counted the number of repeated presentations of each image, including these one-target trials. To plot accuracy as a function of the number of repeated stimulus presentations (Fig. 1d inset), we computed the correct rate using all stimuli in the two-target trials at a given number of repetitions across training sessions. The correct rate for each image (Fig. 1e, Supplementary Fig. 4) was calculated using the sessions after the monkey performance reached around 75% correct (the last 3-5 sessions). The generalization performances for novel images and

manipulated images (Figs. 1i, 2b-d, 4b, Supplementary Figs. 3d, 5) were all calculated using only the first trial of each stimulus.

Reaction times were calculated as the time from stimulus onset to reaching a target (Figs. 1f, 3b, 3d, Supplementary Fig. 9). Thus, the time included both decision formation, motor planning, and execution. Nonetheless, we used this definition because the speed of reaching could also be strongly influenced by the decision-making process (Kaufman et al. 2015). We also confirmed that the time between stimulus onset and reach onset seemed to reflect stimulus difficulty, although the difference in this time was small across images. During reaching, monkeys occasionally used multiple swiping movement to reach a target (12 % of trials), but the results did not qualitatively change even after removing these trials.

Construction of dissimilarity matrix and dendrogram from one-versus-other tasks (Figure 3k-l)

Using eight one-versus-other tasks (Fig. 3i; see the "one-versus-other tasks" section above), we derived the hierarchical organization of category similarity. For each task, we first calculated the proportion of error trials for each of the seven distractor categories, in which the monkey erroneously chose the target box associated with the target category of the task. This was calculated using the last 3-4 sessions in which the monkey's overall correct rate had plateaued (Fig. 3j). Combining the results of all the tasks yielded an 8×8 asymmetric confusion matrix, where the row was each task and the column was each confused category (Supplementary Fig. 11a). We noticed that the previous task affected the current task in this confusion matrix; when the monkey performed the A-versus-other task before the B-versus-other task, they tended to continue to choose A as a target in the B-versus-other task. We therefore removed the target category of the preceding task from each row (Supplementary Fig. 11a). Finally, we symmetrized the confusion matrix by averaging the confusion rates for each category pair from the original non-symmetric matrix. From the matrix, we generated a dendrogram using *linkage* function in the Python *scipy* package (Fig. 3l).

Analysis of multi-object arrangement task for humans (Figure 3l)

To estimate the category hierarchy structure in humans using the multi-object arrangement task (Fig. 3l, Supplementary Fig. 11b), we followed Kriegeskorte et al. (2012) and generated a representational dissimilarity matrix (RDM). In this task, participants arranged object images such that similar images were placed nearby. From the arranged object positions, we calculated the distance between each pair of images. There were multiple images belonging to one category (e.g., two monkey images) in each trial, and we averaged the positions of these images before calculating the distance between the category pairs. We then derived an RDM by combining the results of all trials, following the "Weighted average of iteratively scaled-to-match subset dissimilarity matrices" in Kriegeskorte et al. (2012). In brief, we started with a random RDM, scaled the sum of the distances of all pairs in each trial to match that of the RDM, and then replaced each element of the RDM with the scaled distance of the corresponding pair in that trial. We repeated this procedure

for all trials while giving more weight to trials with greater distances for a pair. This iterative adjustment was repeated until the RDM converged. This procedure was performed on the data from individual human participants, and the resulting RDMs were averaged across participants.

Model comparison

Category classification and behavioral fitting procedures (Figure 2e)

We evaluated the performance of various vision models in classifying object concepts and fitting monkey behavior by constructing linear classifiers based on these model outputs (Fig. 2e, Supplementary Figs. 6-8). The outputs of many models are highly overparameterized ($10^2 - 10^5$ parameters), whereas the number of images we used in each task was limited ($10^2 - 10^4$). Therefore, we first performed a dimension reduction on the model outputs and then employed a cross-validation approach to evaluate their performance. The cross validation was performed by constructing a classifier using the training set of each task and calculating the model's correct rate using the generalization set, mirroring the monkey experimental procedure. For dimension reduction, we performed principal component analysis (PCA) on the model outputs of the training set and selected the first 20 PCs. As explained later, the number of dimensions did not affect our conclusions. We then normalized each dimension and constructed a linear classifier of object concepts or monkey choices using a logistic regression. Finally, we projected the model outputs for the generalization set on the same PC space and calculated the proportion of stimuli correctly classified by the trained linear classifier.

A linear classifier was constructed by fitting the training data to the following linear logistic function:

$$logit[P_i(category1)] = M_i \cdot w + b \tag{1}$$

where $P_i(category1)$ is the probability of classifying image i into category 1, M_i is the dimension-reduced model output for image i, w is the linear weight, and b is a bias parameter. We then computed $P_i(category1)$ for each of the generalization images and calculated the correct rate assuming that the model chooses category 1 if $P_i(category1)$ is greater than 0.5. For the fitting to monkey behavior, we again performed the linear logistic regression but estimated w and b to maximize the fit performance to monkey choices in the training trials of the sessions after reaching a plateau performance. The fit performance was quantified as the proportion of generalization images for which the monkey and model choices matched. In the figures, we showed the fit results aggregated across the three monkeys.

We confirmed that changing the details of this model evaluation procedure did not qualitatively affect our results. First, we tested different numbers of dimensions being reduced by PCA (Supplementary Fig. 6a). All the models quickly improved their cross-validated performance as the number of dimensions increased, and their performance reached a plateau before or near 20 dimensions. We therefore chose this dimension in the main figure. Second, we tested another dimension reduction technique. We constructed a representational dissimilarity matrix of the images by computing Pearson's correlation of the model outputs for each image pair. We then performed a

classical multi-dimensional scaling (MDS) to obtain a 20-dimensional space capturing the similarity pattern. The classification based on these 20 dimensions was similar to the main results (Supplementary Fig. 6b). Third, we replaced logistic regression with a linear support vector machine (SVM) and found that the classification accuracy did not substantially change (Supplementary Fig. 6c).

Model details

V1-like model We followed Pinto et al. (2008) to create a population of V1-like responses, which is generated from a set of Gabor filters of different orientations and spatial frequencies with normalization and non-linear thresholding. In brief, we first calculated the z score of the pixel intensity of an input image (adjusted to 150 $\times$ 150 pixels) and normalized its local part by subtracting the mean and dividing its Euclidean norm. We then convolved it with Gabor filters (43 $\times$ 43 pixels) spanning 16 orientations and six spatial frequencies (1/2, 1/3, 1/4, 1/6, 1/11, 1/18 cycles/pixel) with a fixed Gaussian envelope (standard deviation of 9 cycles/pixel in both directions) and fixed phase (0) (96 filters in total). We then applied an output non-linearity by setting all negative values to zero and clipping the values at one. Finally, we applied divisive normalization to each filter output.

Luminance+color model The model computes marginal statistics of color and luminance information at both the global and local levels. Each pixel intensity of an input image was converted into luminance (cd/m^2) and CIE-u'v' coordinate via the calibration data of the touchscreen monitors obtained using a spectrometer (X-Rite i1 Pro, X-Rite, Grand Rapids, MI, USA). At the global level, the mean, standard deviation, skewness, and kurtosis of the distributions of luminance, u', and v' were calculated for the entire image. At the local level, we segmented the image into 4 $\times$ 4 blocks and calculated these statistics for each segment. All of these values were normalized and then concatenated as the output of the model.

Gist+SF model The Gist model, adopted from Oliva et al. (2006), captures the spatial structure of an image by computing the orientation and spatial-frequency powers of each local image region. Although the extracted features are similar to those of the V1-like model, this model captures larger-scale image structures with more compact representations. In brief, each image was divided into 4 $\times$ 4 segments and oriented Gabor filters (8 orientations) were applied over different scales (4 scales) in each segment. The average filter energy was subsequently calculated for each filter and segment, and all outputs were concatenated. To capture global image patterns, we also calculated the spatial frequency and orientation power (4 frequencies and 8 orientations) of the whole image without segmentation and concatenated them with the outputs. The images were converted to grayscale before the calculation and all the statistics were normalized before being concatenated.

Texture model The texture model consists of statistics that capture the correlation structures of V1-level filter outputs across different scales, orientations, and positions (Portilla and Simoncelli 2000). These statistics are sufficient to recover naturalistic textures (Portilla and Simoncelli 2000) and explain part of the neural responses in macaque V2 and V4 (Freeman et al. 2013; Kim et al. 2019; Okazawa et al. 2015). We adopted the code of Portilla et al. (2000), which used steerable pyramids to compute the outputs of oriented bandpass filters. We chose 4 orientations, 4 scales, and 7 spatial neighborhoods to compute spatial correlations. The texture model also included marginal statistics of the luminance distribution and the powers of the V1-level filter outputs. The images were converted to grayscale prior to the calculation. All the statistics were normalized and then concatenated.

Contour model For the background-cropped image sets, we computed statistics derived from the outline contour of an object image (Supplementary Fig. 8a). The statistics included the aspect ratio (horizontal vs. vertical extent of the contour), the angle of the longitudinal axis (the most elongated axis), the extent of elongation (the ratio between the most elongated axis and its orthogonal axis), the distance of the contour from the image origin at each 5-degree angle, the standard deviation of these distances, and the spatial frequency powers of the contour shape at 4 scales. Thus, these parameters quantified information such as to how much and in which direction the object contour is elongated and how smooth the contour is, as well as the coarsely sampled location of the contour itself.

Deep neural network (DNN) models We used five DNN models commonly used in vision science research —AlexNet (Krizhevsky et al. 2012), VGG-16 (Simonyan and Zisserman 2014), ResNet-50 (He et al. 2015), Vision Transformer (ViT-B/32; Dosovitskiy et al. 2020) and CLIP (Radford et al. 2021). AlexNet, VGG16, and ResNet-50 are convolutional neural networks (CNNs) with different numbers of convolutional layers and network architectures. We used these networks pre-trained on 1,000 object categories in the ImageNet1k dataset. Using the hook mechanism in PyTorch (version 2.0.1), we extracted the activity of the penultimate layer for each image as model outputs. Vision Transformer was used to demonstrate how a DNN with a very different architecture from CNNs performs in our tasks. We extracted the activity of the "encoder" block of ViT-B/32 for each image. CLIP is a recent DNN that learns visual representations by aligning feature spaces between image and text data (Radford et al. 2021). We included this model to demonstrate whether a language-informed model performs differently in our tasks. We chose CLIP with ViT-B/32 as the architecture of visual processing and extracted the activity of its visual encoder for each image.

Calculation of the geometric signal-to-noise ratio (SNR)

To quantify the degree to which object concepts can be separated from the output of the visual models (Fig. 4e), we calculated the geometric SNR, which is the distance between two clusters of images (manifolds) corresponding to two concepts based on the manifolds' geometric properties (Sorscher et al. 2022; St-Yves et al. 2024). A full expression, a detailed explanation, and imple-

mentation can be found in Sorscher et al. (2022). In short, the geometric SNR for concept manifold a when it is classified from concept b is calculated as

$$\text{SNR}_a = \frac{1}{2}\left(\frac{\|\Delta \mathbf{x}_0\|^2 + (R_a^2 R_b^{-2} - 1)/m}{\sqrt{D_a^{-1}/m + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_b\|^2/m + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2}}\right) \tag{2}$$

where $\|\Delta \mathbf{x}_0\|^2$ represents the distance between the centroids of the two manifolds in the output space, R_i^2 is the mean squared radius of the manifold i, D_a is the "participation ratio" that quantifies the number of dimensions along which the manifold a varies significantly, and $\|\Delta \mathbf{x}_0 \cdot \mathbf{U}_i\|^2$ quantifies the overlap between the signal direction $\Delta \mathbf{x}_0$ and the manifold axes of variation $\mathbf{U}_i$. m is the number of stimuli used to train the classifier, and we used $m = 5$ for Fig. 4e as a demonstration but confirmed that the relative magnitudes across models were largely stable over a range of m. Since the geometric SNR is an asymmetric quantity, we took the average of SNR_a and SNR_b. We also tested another conventional metric (silhouette score) to quantify the degree of clustering and found qualitatively similar results.

Comparison of monkey and human behavior (Figure 3)

We compared monkey and human behavior using the grayscale training image sets (Fig. 1b, g) because these images were repeatedly shown to the monkeys; thus, their performance on individual images could be measured. Monkeys exhibited variable choice accuracy across images (Fig. 1e, Supplementary Fig. 4), whereas humans showed near-perfect performance (Fig. 3a) with different reaction times (RTs) (Fig. 3b). We therefore developed a metric to quantify the performance for each image using both choice accuracy and RTs. We took two approaches and confirmed that both approaches yielded similar results. In the first approach (main figures), we leveraged the drift-diffusion model that has been used to model choices and RTs in perceptual decision making (Okazawa and Kiani 2023). In the second approach (Supplementary Fig. 10), we followed previous literature that used a simpler metric derived from correct rates and RTs as explained below.

Metric based on the drift-diffusion model (DDM)

To quantify the difficulty of categorizing each stimulus, we used DDM, which assumes a common underlying mechanism for choices and RTs. Many previous studies have used DDM to propose that the brain accumulates sensory evidence over time, but in this work, we did not have a strong assumption that evidence accumulation underlies behavior in our task (Stine et al. 2020). Rather, we used DDM only to derive a metric of stimulus difficulty that could be compared between the human and monkey results. We believe that the high fitting accuracy of this model (Fig. 3d, Supplementary Fig. 9) justifies the use of such a metric.

Our key assumption is that the drift rate (category sensitivity) varies across images and reflects stimulus difficulty, whereas the decision bound and non-decision time are stable across images for

each participant and task. For each image, the drift-diffusion process forms the decision variable at time t ($v(t)$) as

$$v(t) = \int_0^t dr_i + \eta(\tau)\, d\tau \tag{3}$$

where dr_i is the drift rate (category sensitivity) for image i. $\eta(\tau)$ represents a noise term following a Gaussian distribution with a mean of 0 and an SD of 1. When $v(t)$ reaches a positive ($+B$) or negative ($-B$) bound, the model commits to the decision associated with the bound (category 1 or 2). The reaction times are modeled as the time required to reach a bound plus an additional delay (non-decision time) common to all images. We modeled the non-decision time as following a Gaussian distribution with a mean of T_0 and an SD of $T_0/3$.

The model had two free parameters (B, $T0$) common to all images and the drift rate (category sensitivity; dr_i in Eq. 3) as a free parameter for each image (total $n + 2$ free parameters; n is the number of images). Higher drift rates indicate that images were easier to classify. We fitted the model to each participant's choices and RT distributions in each task using maximum likelihood estimation (Luo et al. 2025). Given a set of parameters, the RT distributions of the two choices were calculated according to the aforementioned model formulation by numerically solving a Fokker-Planck equation. From these distributions, we derived the likelihood of observing the participants' choices and RTs. Summing the likelihoods over all trials yielded the total likelihood of the parameter set. We used a simplex search method (*fminsearch* in MATLAB) to determine the parameter set that maximized the summed likelihood. As shown in Supplementary Fig. 9, our model fitted both monkey and human RTs and correct rates accurately, ensuring the use of the drift rate as a metric of stimulus difficulty. Figure 3e-h showed the drift rate averaged across participants.

Inverse efficiency metric

As an alternative to the model-based approach to estimating stimulus difficulty, we used a modified version of a simple metric called "inverse efficiency" (Koopman et al. 2017), which quantifies task performance using correct rates and RTs as follows:

$$I = \frac{medianRT}{P(correct) + 1} \tag{4}$$

Here, the median RT was calculated for each image after outliers ($\pm 2SD$, within each image) were excluded. It was then divided by the correct rate for that image. The denominator included 1 to avoid extreme value when the correct rate was low. This metric was calculated for each participant and then averaged across participants (Supplementary Fig. 10).

Supplementary figures

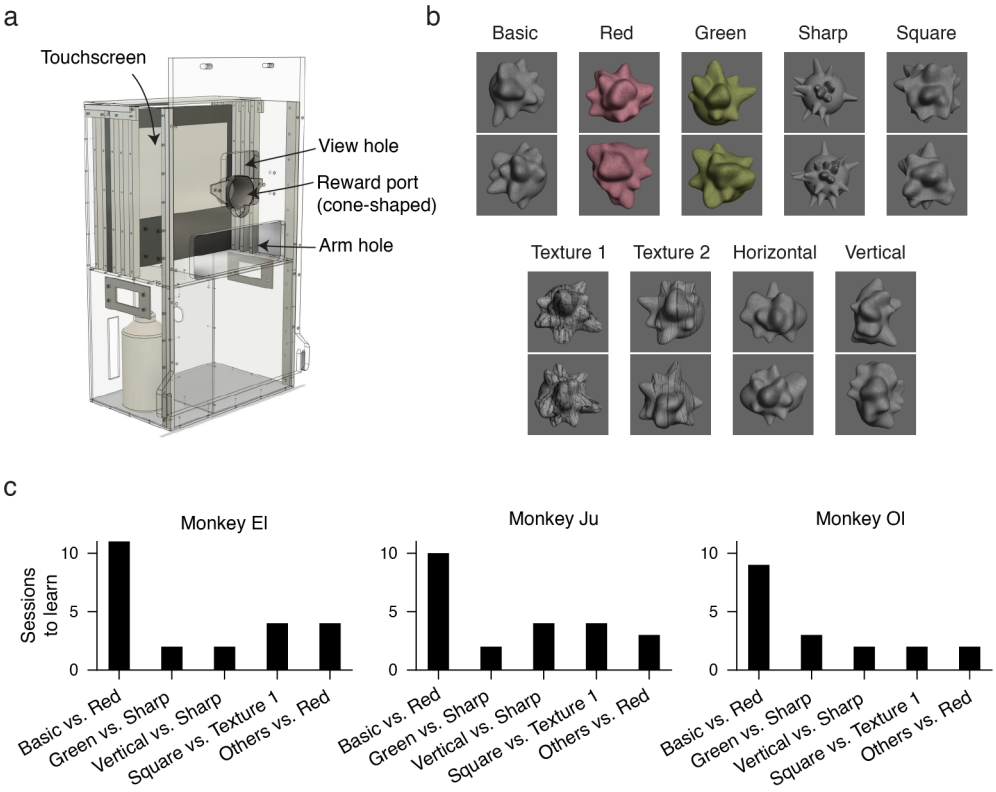

Figure S1: ***Pretraining of monkeys on the task structure using shape stimuli.*** *(**a**) Behavioral testing system attached to the monkeys' home cage. Monkeys reached for a touchscreen through an arm hole in the front panel and received a juice reward from a cone-shaped reward socket (Kawaguchi et al. 2019). While performing tasks, they placed their mouth on the reward socket and looked at the screen through a viewing hole. (**b**) Prior to the main concept classification tasks, we trained the monkeys to perform the classification of amorphous shape stimuli. The task sequence followed the main design (Fig. 1a), but the monkeys had to categorize shape stimuli according to their shape, color, or texture. We created nine groups of stimuli as shown in this panel and chose two of them to classify for each task. In each group, there were 50 images with random bump positions. (**c**) The number of sessions (days) required for the monkeys to learn each shape classification task plotted in the trained order. Prior to the first task, monkeys were trained to touch a shape stimulus and move it to a target box. Then, in the first "Basic" vs. "Red" task, they were presented with two target boxes and had to learn to select one of them based on a stimulus feature. This initial stimulus-response association training took 9-11 days (see Methods). The monkeys were subsequently trained to classify different pairs of shapes in 2-5 days each. In the "Others" vs. "Red" task, the monkeys had to choose one target for red stimuli and the other target for stimuli randomly selected from the remaining eight stimulus categories.*

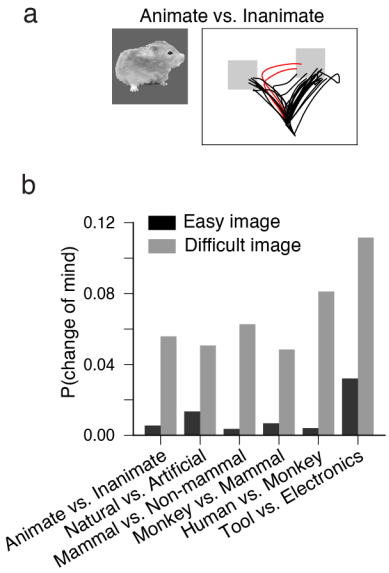

Figure S2: ***Monkey reach trajectories reflected their decision-making process.*** *(**a**) Monkeys moved a stimulus to a target with a straight path in most trials, but they occasionally made detours as if they changed their mind (red lines; Kaufman et al. 2015). These trials were detected by looking for trajectories whose average position was the opposite to the position of the chosen target with respect to the midline. The right target box corresponded to the animate category in this example. (**b**) Monkeys exhibited these detour trajectories much more often for difficult images. For each task, we selected the 10 easiest and most difficult images based on correct rates and plotted the probability of trials in which the monkeys exhibited detour trajectories. The plots aggregate the data from the three monkeys.*

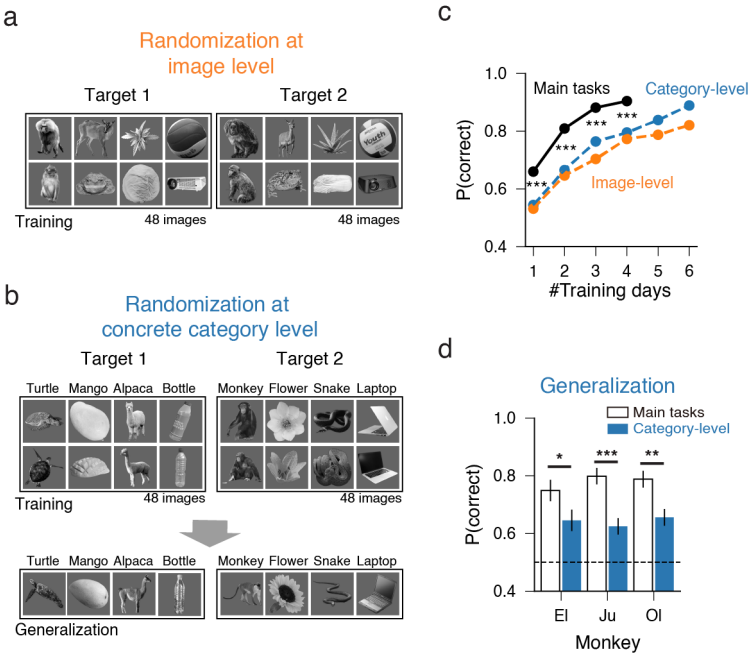

Figure S3: ***Absence of a common conceptual rule results in poorer performance in stimulus-response associations.*** *(**a-b**) We assessed how well monkeys could learn stimulus-response associations in the absence of shared conceptual rules for classifying images using two control tasks. In one task (a), two targets were associated with randomly selected images from our grayscale cropped image set (48 images per target). However, the same image was always associated with the same target, allowing the monkeys to improve their performance by remembering individual associations. In the other task (b), two targets were associated with random concrete object concepts (e.g., apple, horse, bin; referred to here as the "concrete category"; 16 concrete categories for each target). Two to six images within the same concrete categories were always associated with the same targets. During the generalization test, different images of the same concrete categories were shown. (**c**) Although the monkeys could gradually learn both versions of the randomization tasks, the performance during learning was consistently lower than that on the main concept classification tasks ($p < 10^{-10}$ for all days aggregated across monkeys, binomial test). For the main tasks, we calculated the average of the six tasks shown in Fig. 1. (**d**) After training on the concrete category randomized task (b) for 6 days, we tested their generalization performance on new stimuli. Their performance was lower than the generalization performance of the six main tasks ($p < 0.016$ for all three monkeys, binomial test). Error bars are S.E.M. across images. Asterisks indicate statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.*

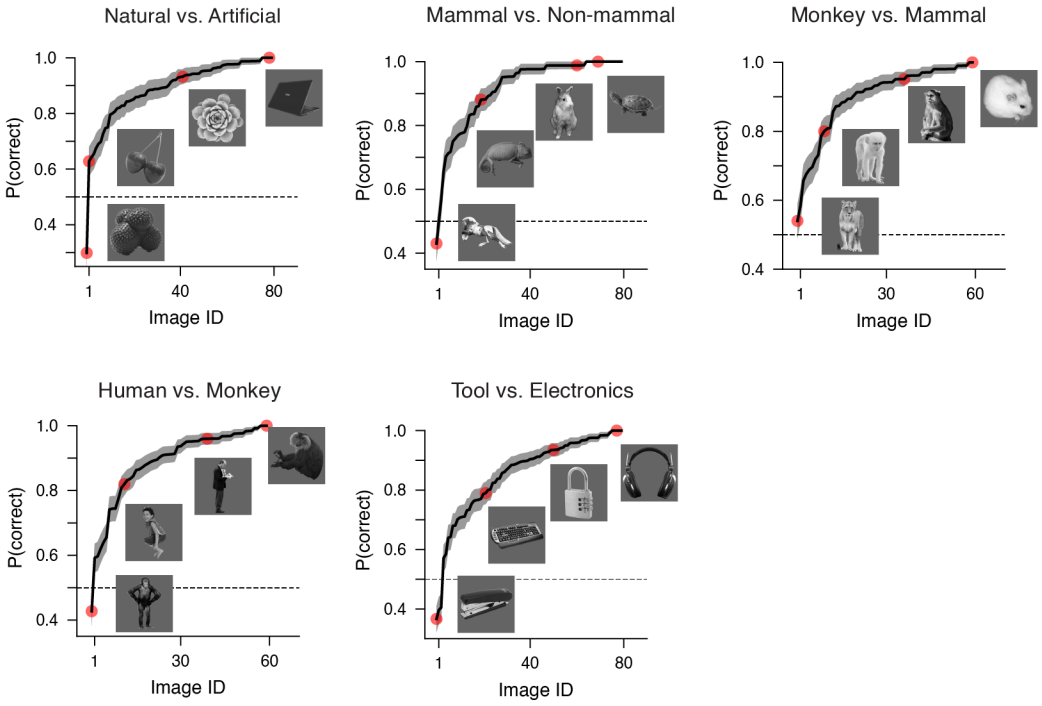

Figure S4: ***Correct rates varied substantially across images in each task.*** *In Fig. 1e, we showed that the monkeys performed well on many images and poorly on some other images in the animate vs. inanimate task. The same was the case for the other five concept tasks in Fig. 1 ($\chi^2 > 259$, $p < 10^{-10}$ for all tasks, χ^2-test). Each panel shows the average performance of the three monkeys for each task, excluding the first 1-2 training sessions in which the overall performance did not reach $\sim$75% correct. We found that the monkeys tended to incorrectly categorize images that also looked confusing to humans. For example, in the natural vs. artificial task, they performed poorly on an image of a rubber toy that also looked like a fruit. In the human vs. monkey task, an image of a standing ape was often classified incorrectly. The shading indicates S.E.M. across trials.*

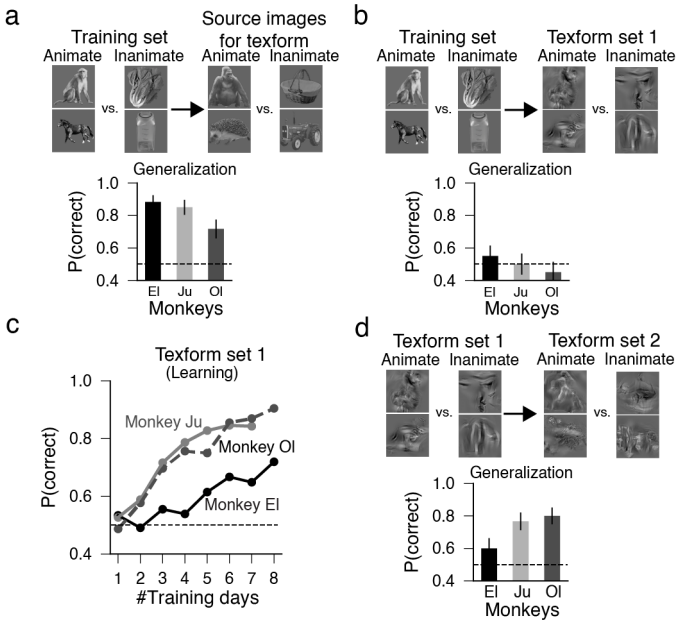

Figure S5: ***Monkeys failed to generalize a rule to texform images.*** *It has been shown that animacy information is partially preserved even in images that retain only mid-level textural visual features (texform; Long et al. 2018). We found that the monkeys failed to generalize the learned animate vs. inanimate rule to texform images. Thus, it is unlikely that the monkeys relied primarily on mid-level textural features in our main task. Error bars are S.E.M. across images. (a) We first confirmed that monkeys could generalize the rule to the original, source images (object photographs) used to generate texform images by Long et al. (2018). (b) However, when we showed texform images, the monkey performance decreased to the chance level ($p = 0.26 - 0.82$, $BF_{01} = 4.7 - 6.3$, binomial test). We used 60 texform images (set 1). (c) We then attempted to train the monkeys to discriminate animate vs. inanimate using the same set of texform images (Set 1). Although the monkeys gradually improved their performance was consistently lower than that of the main concept tasks for all monkeys ($p =< 9.5 \times 10^{-4}$ for the first 4 days). In particular, monkey El presented a much slower learning curve. (d) Monkeys Ju and Ol were able to generalize the learned classification to other texform images ($p < 2.1 \times 10^{-5}$), but the performance of monkey El was statistically indistinguishable from chance level ($p = 0.078$, $BF_{01} = 1.9$, binomial test).*

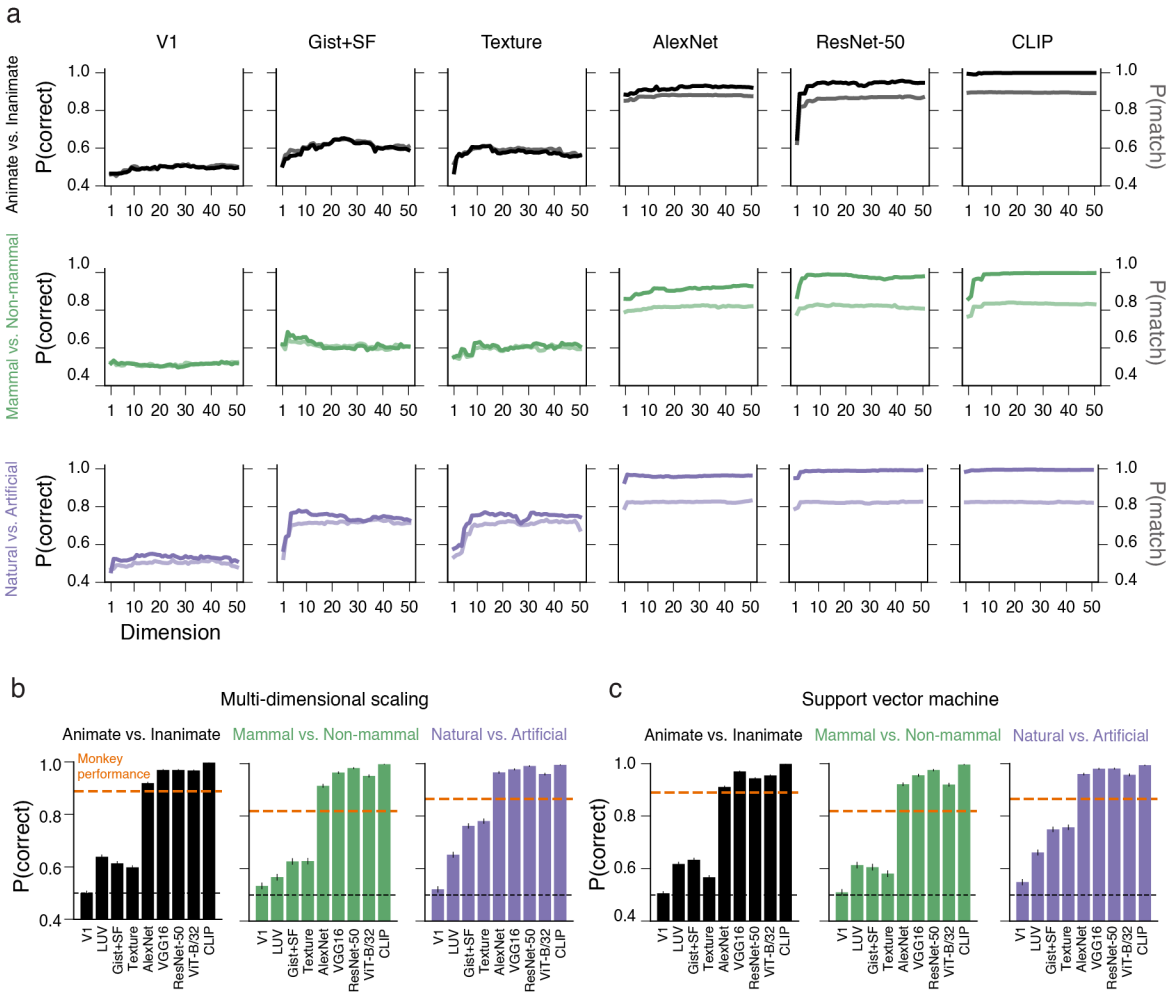

Figure S6: ***Model performance did not depend on classification or dimension reduction methods*** (***a***) *We confirmed that the classification and behavioral fitting performances of the models (Fig. 2e) did not strongly depend on the number of dimensions of the model parameters reduced by principal component analysis (PCA). In each panel, the darker line represents the model performance in correctly classifying object concepts, and the lighter line represents the model performance in fitting monkey choices. In both cases, the fitting was performed using the training image sets, and the performance was evaluated using the generalization image sets used in the monkey experiments shown in Fig. 2.* (***b, c***) *Model performance was not strongly dependent on the methods used for dimension reduction and classification. We tested multi-dimensional scaling for dimension reduction (b) and used a support vector machine instead of logistic regression for classification (c) and found that the results did not change qualitatively.*

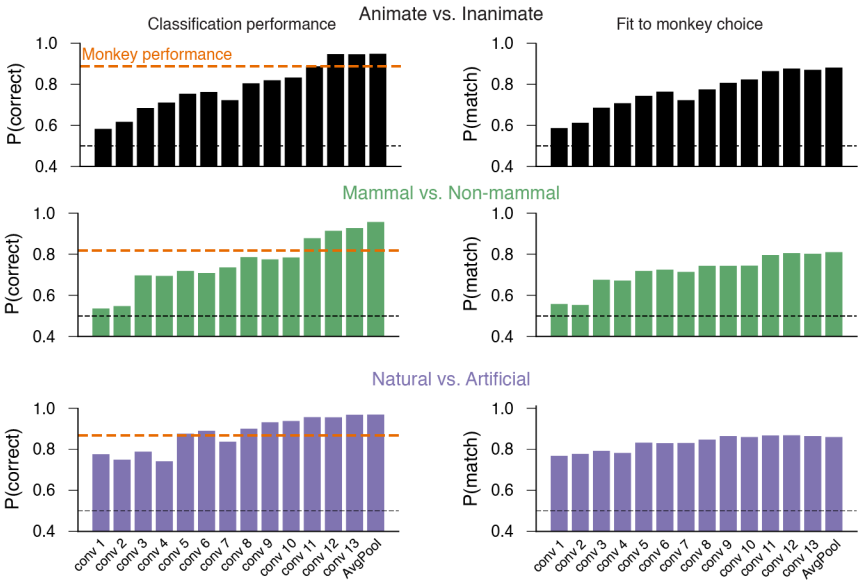

Figure S7: ***Deeper layers of deep neural networks (DNNs) better classify abstract object concepts and fit monkey choices.*** *We confirmed that the deeper the layers of the DNNs are, the better they perform at classifying abstract object concepts (left) and fitting monkey choices (right). The panel shows the results from VGG16. We also observed similar results for other DNNs. The conventions are the same as those in Fig. 2e.*

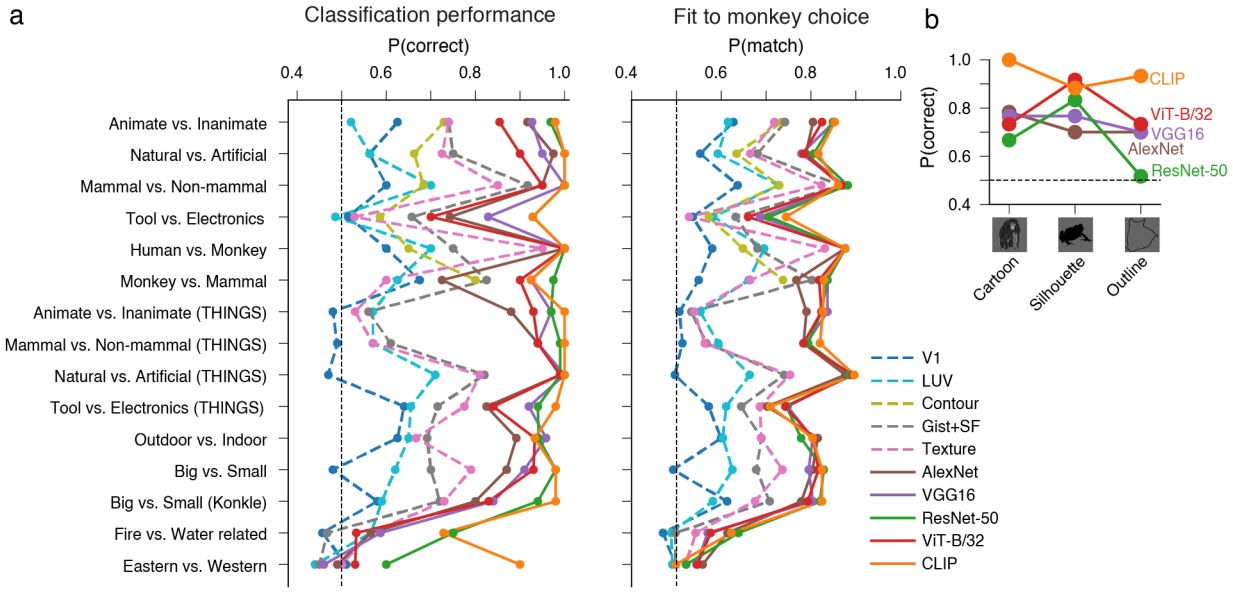

Figure S8: ***Model classification and behavioral fitting performance for all tasks.*** *(a) Summary of model performance for all tasks shown in Figs. 1, 2, and 4. For the tasks using large-scale THINGS images (Fig. 2), we used only the first 100 generalization stimuli to compute the performance for proper comparison; thus, the performances do not exactly match those shown in Fig. 2e. The classification performances of ResNet-50 and CLIP are the same as those shown in Fig. 4b. (b) DNNs could also classify modified stimuli outside the range of natural images (cartoon, silhouette, and outline images), which we tested with monkeys (Fig. 3g). The performance tended to be lower for the outline images, which is consistent with the monkey and human behavioral results (Fig. 3g).*

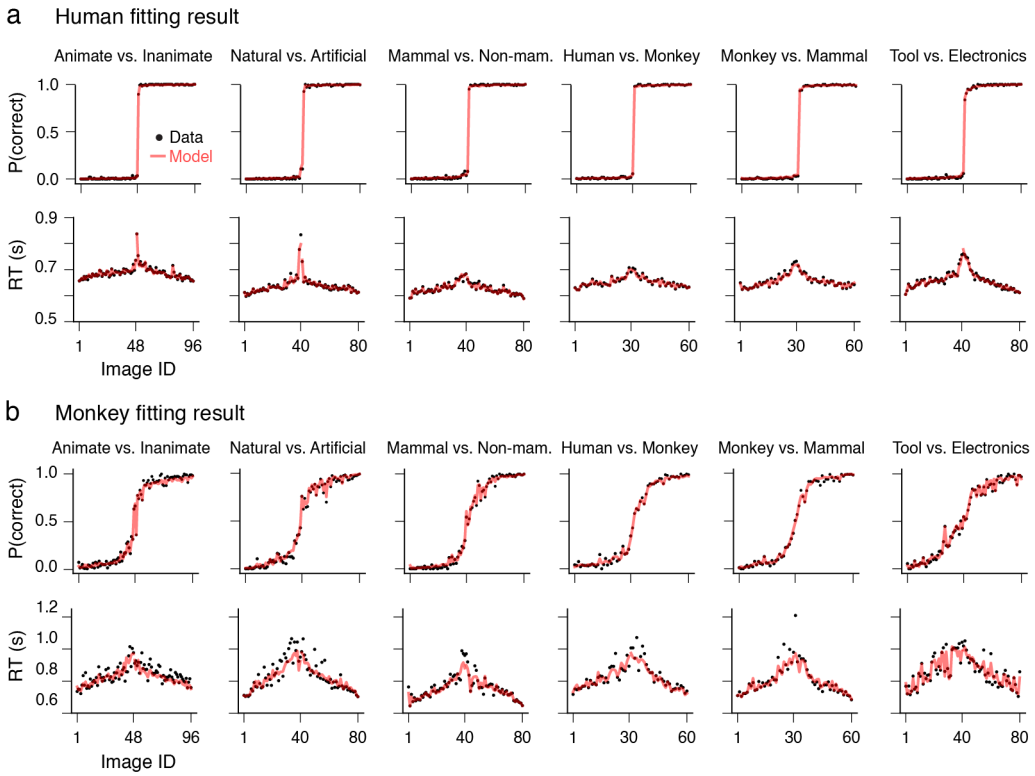

Figure S9: ***Drift diffusion model (DDM) could accurately fit choices and reaction times.*** *(**a, b**) Fitting of the DDM (Fig. 3c) to the behavioral data of individual tasks for humans (a) and monkeys (b). The conventions are the same as those in Fig. 3d. Model fit was performed for the tasks using grayscale image sets (Fig. 1), in which the monkeys repeated many trials for the same stimuli. See Methods for the details of the model fit. The image IDs were sorted by the sensitivity parameter (dr_i in Eq. 3) of the model fit averaged across participants for each task. The fit lines are not smooth because the plots are the average across participants who had different rank orders of sensitivity across image IDs.*

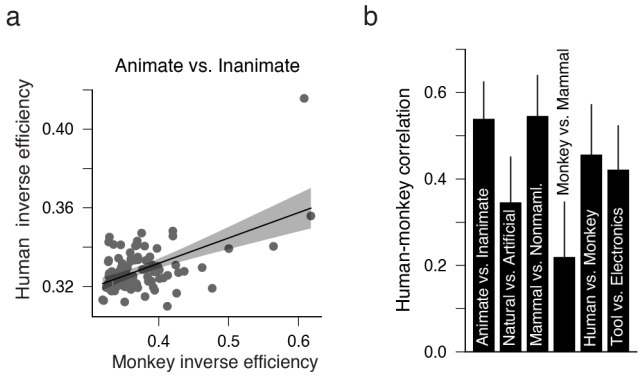

Figure S10: ***Moderate correlations between human and monkey behavior could be confirmed using a different metric.*** *(**a**) As an alternative to the drift-diffusion model to quantify and compare performance between humans and monkeys (Fig. 3), we used a simple metric that divided the median reaction time by the probability of correct choice ("inverse efficiency"; Eq. 4). We calculated this metric for each image and tested the correlation of this metric between humans and monkeys in each task. The panel shows the results of the animate vs. inanimate task as an example. (**b**) This metric also showed moderate correlations between human and monkey behavior, which was consistent with the main result (Fig. 3f) except for the monkey vs. mammal task. Error bars are S.E.M. across images.*

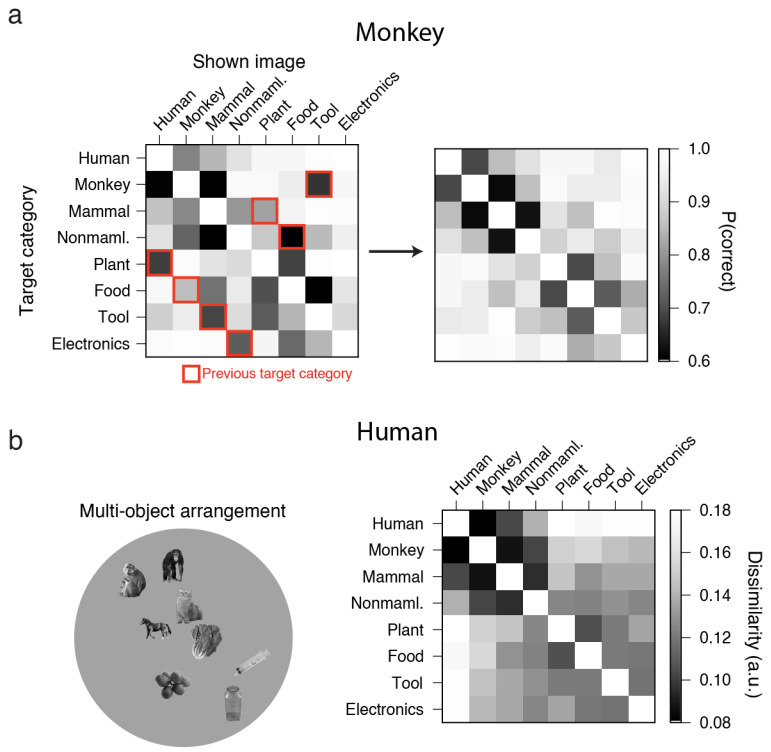

Figure S11: ***Construction of a similarity matrix for eight object groups.*** *(**a**) Using monkey behavioral data from one-versus-other tasks (Fig. 3i), we constructed a confusion matrix for eight object groups. We first calculated the correct rate for each object group (column) in each one-versus-other task (row) (left panel). For example, the correct rate in the third column and the first row indicates the probability that the monkey correctly classified mammal images into the "other" target in the human-versus-other task. We then performed two steps of post-processing on this matrix (right panel). First, we noticed that the monkey showed a strong history effect and tended to make large errors for the object group that was the target in the last task (red squares). Therefore, we removed these data points from the matrix. Second, we symmetrized the similarity matrix by averaging the correct rates of the two matrix elements for each object pair. (**b**) To generate a similarity matrix for humans, we performed a multi-object arrangement task, in which participants ($n = 6$) arranged multiple object images on a screen such that the distance between each image pair reflected the perceived similarity of these objects (left panel; Kriegeskorte and Mur 2012). The similarity matrix was generated using an optimization approach that iteratively searched for the matrix that best accounted for the participants' image placement across all trials (see Methods). The plot is the average of the matrix across the participants.*

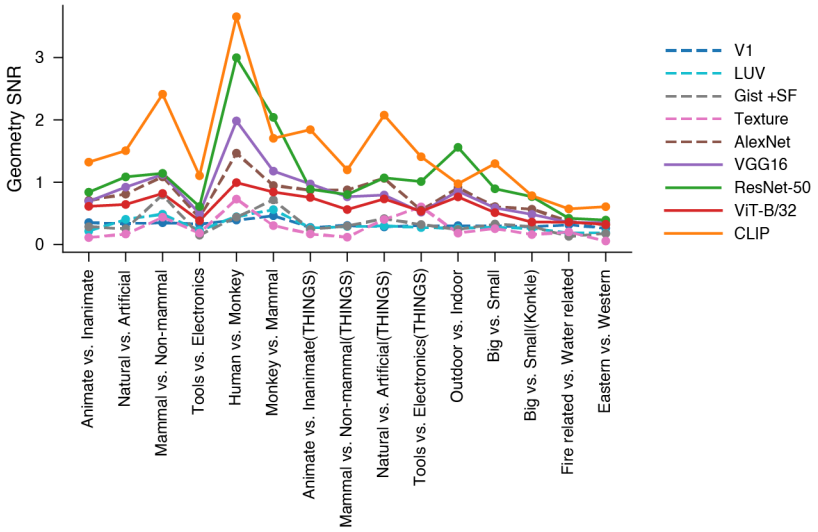

Figure S12: ***Geometric signal-to-noise ratios (SNRs) of all the models for all the tasks.*** *The degree of clustering of object concepts in the outputs of visual models for each task is measured as the geometric SNR (Sorscher et al. 2022). Fig. 4e shows the average of these values across the tasks that the monkeys were able to perform (i.e., the tasks other than fire- vs. water-related objects and Western- vs. Eastern-culture objects). For most tasks, DNNs showed greater concept separation than low- and mid-level visual models did, with CLIP showing the greatest separation.*

Movies 1-3

Example videos showing the monkeys' performance on the tasks using the large-scale THINGS database (Fig. 2; Movie 1: monkey Ju performing the animate vs. inanimate task, Movie 2: monkey El performing the natural vs. artificial task, Movie 3: monkey Ol performing the mammal vs. non-mammal task). The videos were taken during the generalization session, where half of the images were from the generalization set, which was shown only once in each task. The other half were from the training set. The upper parts of the videos are screenshots of the monitor the monkey was looking at.