Perceptual Category Learning of Photographic and Painterly Stimuli in Rhesus Macaques (*Macaca mulatta*) and Humans

Drew Altschul¹², Greg Jensen³, and Herbert S. Terrace³

¹University of Edinburgh  
²Scottish Primate Research Group  
³Columbia University

ABSTRACT

Humans are highly adept at categorizing visual stimuli, but studies of human categorization are typically validated by verbal reports. This makes it difficult to perform comparative studies of categorization using non-human animals. Interpretation of comparative studies is further complicated by the possibility that animal performance may merely reflect reinforcement learning, whereby discrete features act as discriminative cues for categorization. To assess and compare how humans and monkeys classified visual stimuli, we trained 7 rhesus macaques and 41 human volunteers to respond, in a specific order, to four simultaneously presented stimuli at a time, each belonging to a different perceptual category. These exemplars were drawn at random from large banks of images, such that the stimuli presented changed on every trial. Subjects nevertheless identified and ordered these changing stimuli correctly. Three monkeys learned to order naturalistic photographs; four others, close-up sections of paintings with distinctive styles. Humans learned to order both types of stimuli. All subjects classified stimuli at levels substantially greater than that predicted by chance or by feature-driven learning alone, even when stimuli changed one every trial. However, humans more closely resembled monkeys when classifying the more abstract painting stimuli than the photographic stimuli. This points to a common classification strategy in both species, on which humans can rely on in the absence of linguistic labels for categories.

Keywords: cognition, categories, simultaneous chain, rhesus macaques

INTRODUCTION

When a human perceives a stimulus, she will automatically identify that stimulus as belonging to one or more categories. We can describe this process linguistically: when we perceive something, we do not simply see the thing; we see the something as a type of thing, part of a set of things which can be in some way treated as equivalent (Murphy et al., 2012). Categorization of visual stimuli occurs so rapidly in humans that it feels almost instantaneous (Grill-Spector and Kanwisher, 2005), and we interpret the entire world through a diverse set of classifications (Goldstone and Kersten, 2003).

Do non-human animals (hereafter, “animals”) possess the same categorization abilities as humans? This question has motivated considerable research by comparative psychologists, who have sought to understand evolutionary roots of categorization through the study of many species. The resulting literature has demonstrated that animals have the ability to categorize a bewildering range of stimuli, including organic forms, such as faces (Marsh and MacDonald, 2008), plants, and animals (Roberts and Mazmanian, 1996; Vonk and MacDonald, 2002; Vonk, 2013), as well as man-made objects, such as cars, chairs (Bhatt et al., 1988), orthographic characters (Schrier et al., 1984), paintings (Watanabe, 2013), cartoons...
Matsukawa et al., 2001), and abstract forms (Vogels, 1999). Animals can also correctly categorize never-before-seen exemplars, showing that this ability is not limited to previously-learned stimuli (Schrier and Brady, 1987; Sigala, 2009). Perhaps most impressively, animals can categorize some images based only on exposures lasting less than 100ms (Richard and Thorpe, 1998), suggesting that animals can process image features in parallel, rather than performing a systematic visual search for feature cues. These sophisticated abilities have been reviewed extensively elsewhere (Jitsumori and Delius, 2001; Miller et al., 2003; Katz et al., 2007; Zentall et al., 2008).

Researchers of human cognition distinguish categories from concepts: a category is most often defined as a set of entities which exists in the real world and are grouped together, whereas a concept has a more abstract character. It is generally agreed that a “concept” encodes the mental representation used in human categorization (Goldstone and Kersten, 2003). Researchers take for granted that humans possess countless internal concepts, but no such consensus exists for whether animals make use of similarly abstract frameworks. Thus, although animals can clearly categorize stimuli, studying the underlying mechanisms is complicated by the difficulty in disentangling the cognitive representations employed by animal minds from other unrelated features of behavior.

Some authors have defined concepts as necessarily linguistic a priori, ruling out mechanisms for non-linguistic concept formation (Chater and Heyes, 1994), and thereby, nearly all potential models for animal concepts. Other authors have argued that animals learn to categorize using reinforcement learning and associative conditioning alone (Roberts, 1996). This associative account of an animal’s classification of similar stimuli (as described by Herrnstein et al., 1976) relies on the associative strength of discriminable features common across a category (Lea, 1984). In the absence of verbal communication with their subjects, comparative psychologists have largely limited their inquiries to studies of rule-based concepts (Close et al., 2010; Wright and Lickteig, 2010). While an animal can demonstrate that it has learned the rules imposed by researchers, it is virtually impossible to study unsupervised and unrewarded assignment of stimuli into categorical groupings without using language to instruct subjects.

The non-human aptitude for categorization thus falls into a middle ground between associative learning at one extreme and linguistic abstraction at another. Many studies argue that more flexible and sophisticated processes than mere associative learning are required to explain performance (Zentall et al., 2008). However, non-human subjects do not possess language, and thus cannot rely on it as a scaffold for categorical inference. The middle ground between these extremes is not well-defined, and no consensus has emerged regarding the appropriate terminology for identifying or describing these cognitive processes.

Herrnstein (1990) attempted to build a bridge between the linguistic and associative accounts by proposing that animals classify stimuli using open-ended categories. These stimulus clusters would hypothetically rely on the similarities between many learned exemplars, but would fall short of fully-fledged concepts. The criteria for this distinction were vague: a discrimination could be attributed to a concept only if a characteristic other than similarity was used to classify novel exemplars. Skeptics retorted that, because stimuli must necessarily have some features in common (without which they would be unrecognizable), those features must permit categorization on the basis of their similarity (Huber, 2000). Thus, Herrnstein’s proposal left the issue unresolved, and the possibility of feature-driven learning remains a major confound for the study of concept formation in animals.

Herrnstein’s definitions have defined the debate on animal categorization and concept formation, but these are at odds with the most common definitions used among researchers of human cognition. For comparative psychologists, possessing an internal representation that allows an animal to reliably classify stimuli does not immediately indicate that the animal has formed a concept. It is intuitive and well documented that adult humans can flexibly and abstractly manipulate their internal concepts (Hampton, 2003), and can use these concepts to mediate their perceptual categorization (Close et al., 2010). The
majority of studies of adult human cognition rely on language, but without either language or the ability to directly inquire into the nature of an animal’s mental representation, inference is restricted. Additionally, the explanatory power of associative learning in animals is very strong (Wasserman et al., 2015). Due to the absence of language and the possibility of associative learning, comparative psychologists must require a high level of evidentiary rigor to accept the claim that animals can form concepts.

By contrast, it is uncontroversial to assert that animals possess percepts, here defined as integrated sensory representations constructed from both primary sensory data and top-down information processing (Leopold and Logothetis, 1996; Pezaris and Reid, 2007; Sovrano et al., 2010). Because percepts involve the hierarchical integration of information from different levels of processing, they constitute a form of representation that is implicit and statistical instead of being explicit and propositional, even in humans (Gauthier et al., 1997). Both humans and rhesus macaques are adept at learning to categorize percepts by integrating information, even in cases where discrete features are not available (Smith et al., 2010). Macaques can also assign percepts to appropriate experimenter-defined categories (Freedman et al., 2001). An important property of percepts is that they are more than a list of isolated features. Sensory integration allows organisms to evaluate percepts as gestalt-like wholes.

We are therefore interested in the question, “How do subjects judge whether the percept of a current stimulus to belongs to a particular category?” Borrowing a term from the computer vision literature, any algorithm responsible for categorizing a percept is called a classifier citepMeye2008, Pere2009. The definition is functional, in that any algorithm used for categorization may be called a classifier. Ordinarily, the term is used to denote the broad family of strategies used by computer models with varying degrees of success. The methods used by brains to process gestalt sensory information remain an open research topic (Reddy et al., 2010), but these biological processes may also be labeled as classifiers. A rigorous understanding of a subject’s classifier is equivalent to a robust theory of the cognitive processing that underlies categorization. Thus, another way to frame our research question is to ask, “How do the classifiers used by monkeys and humans work, and how do the species differ from one another?”

The challenge in answering this question empirically is to differentiate among the many varieties of classifier an organism might be using. Many of the tasks used to study classification in animals could potentially be solved in many ways. For example, “binary categorization” tasks (e.g. Schrier and Brady, 1987; Matsukawa et al., 2001) present subjects with stimuli that belong to one of two categories (e.g. “Is it a house or a face?”). These tasks can be “solved” if some cue exist that identifies of one of the categories. Should such a cue exist (e.g. the presence of shapes with 90-degree angles), a simple associative classifier would be sufficient to choose correctly. In order to rule out trivial classification strategies, a study must have a sufficient level of task complexity (Jensen and Altschul, 2015). A more complex task is one that requires that a larger number of categories be simultaneously considered as candidates.

Consequently, we designed our test procedure to be much more difficult than those used in past studies of animal categorization by modifying the simultaneous chaining paradigm (or “SimChain,” Terrace, 1984). In classical SimChain, subjects are presented with a set of stimuli, and are required to touch each stimulus in a prescribed order to receive a reward (reviewed in Terrace, 2005). Our adapted procedure, the Category Chain, also requires that subjects touch all four categories in a prescribed order, but it uses different exemplars on each trial. Touching the stimuli in a random order would earn a reward less than 5% of the time. To complete a trial successfully, subjects must simultaneously represent all four stimuli and classify them into their respective categories, making it much more difficult to solve the task using a simple shortcut.

Task complexity is not sufficient, however, if stimuli are trivially discriminable from one another. For example, sorting stimuli by color might entail that many different categories be considered (red, green, etc.), but discriminations would still boil down to measuring simple image statistics. Even if stimuli are
individually complex, small stimulus sets are confounded by the possibility of learning correct responses by rote. If, on the other hand, stimuli are numerous, diverse, and difficult to discriminate, then more sophisticated classifiers are needed to solve the problem. To differentiate between classifiers, a study must also display sufficient stimulus set complexity. A complex stimulus set must have three characteristics: (1) It should consist of hundreds of stimuli per category to undermine memorization; (2) Stimuli within each category must differ from one another in a variety of ways; and (3) Stimuli from different categories must also share similarities.

Our study used two different varieties of image categories. In the “photographic” condition, subjects were presented with photographs drawn from four categories of natural images: birds, cats, flowers, and humans. In the “painting” condition, subjects were presented with small sections of paintings by four artists: Salvador Dalí, Jean-Léon Gérôme, Claude Monet, and Vincent van Gogh. The photographic case represented a naturalistic problem, giving subjects the benefits of discrete features (eyes, wings, etc.). The paintings, on the other hand, represent a more abstract problem: the samples were too small to clearly identify the topic of the painting, and all four painters sampled from the whole color spectrum. Thus, correctly classifying the painting hinged on more global image properties.

By using small portions of paintings as stimuli, we sought to challenge feature-based accounts of categorization (e.g. Vogels, 1999; Marsh and MacDonald, 2008). Individual stimuli within each painting category varied dramatically in terms of their primary image statistics (see the supplemental material for more information). As such, category membership could not be defined in terms of discrete features. Successful categorization of the painting stimuli must instead rely on a gestalt appreciation of the image properties. Although such classification could be described in terms of sophisticated statistical learning, it cannot be explained using models of associative strength.

We trained seven monkeys in the Category Chain task. Three learned to classify photographic stimuli, and four learned to classify painting stimuli. We also tested 41 naïve humans, without instruction or any prior experience, on the same Category Chain task, first with photographic stimuli, and subsequently with paintings. The combination of the Category Chain task with our large stimulus sets satisfied both the task complexity and stimulus set complexity requirements needed to rule out the viability of trivially simple or associative classifiers.

MATERIALS & METHODS

Experimental Task: Category Chain

As in traditional SimChain (Terrace, 2005), Category Chain presents four stimuli on screen at the beginning of each trial, in positions that are randomly assigned. Each trial included one stimulus from each of four categories (e.g., a bird, a cat, a flower, and a human). Subjects had to touch each of these items in a particular order in order to receive a reward. Subjects could continue responding as long as they made no errors. However, the first incorrect touch ended the trial, leading to a time-out period and a random rearrangement of the on-screen stimulus positions on the next trial.

What distinguishes the Category Chain from prior SimChain variants is that the specific stimuli for that trial were selected at random from a large image bank. Thus, the cat presented on trial 2 differed from the cat presented on trial 1, and did not appear again for the remainder of the session. In this respect, the specific stimuli are trial unique, and subjects must learn to classify each stimulus according to its respective category before then selecting the stimuli in the correct order.

Figure 1A gives an example of how Category Chain appears over successive trials. Although four stimuli were always presented, each trial used different images. Three example trials are shown in Figures 1. In the correct trial (Figure 1B), each stimulus is touched once, in the order birds-cats-flowers-people.
Figure 1. The Category Chain procedure. (A) Four consecutive trials of the Category Chain task. Each trial presents one stimulus from each of the four categories, but the specific photographs change and the stimulus positions change randomly from one trial to the next. (B) An example of a correct trial. The dashed lines indicate initial touch, and the arrows indicate subsequent touch (neither dashed lines or arrows were visible to subjects). (C-D) Two examples of incorrect trials. In the first case, the initial touch was to the wrong stimulus, so the trial ended immediately. In the second case, the first two responses were correct but the third was incorrect. (E) Two examples each of the four photographic stimulus categories. (F) Two examples each of the four painting stimulus categories.

On one incorrect trial (Figure 1C), the initial touch is an error, ending the trial immediately. On another (Figure 1D), a later touch is an error, also ending the trial. Figure 1E and 1F present example stimuli from each of the categories used in the study. Further details about the stimulus sets, including distributions of their primary image statistics, are provided in the supplemental material.

Animal Subjects

Seven rhesus macaques (Macaca mulatta) performed the Category Chain task. Subjects were housed at the New York State Psychiatric Institute throughout the study. Treatment conformed with the guidelines set by the U.S. Department of Health and Human Services (National Institute of Health) for the care and use of laboratory animals. Three subjects learned to touch stimuli from the photographic categories in a particular order that differed for each subject (Augustus: Flowers → Cats → People → Birds; Coltrane: Birds → Flowers → Cats → People; Lashley: Cats → Birds → People → Flowers). Four subjects learned to touch stimuli from the painting categories in a particular order (Benedict: Dalí → Gérôme → Monet → van Gogh).
van Gogh; Horatio: van Gogh → Monet → Gérôme → Dalí; Macduff: Monet → Dalí → van Gogh
→ Gérôme; Prospero: Gérôme → van Gogh → Dalí → Monet). For both photographs and paintings, subjects received extensive training on category membership prior to their final training with the Category Chain task. This prior training is described in the supplemental material. Following training, all sessions consisted of 40 trials, each of which consisted of a varying number of responses. Subjects had to respond to each item in the correct order to earn a reward.

The housing, operant chambers, and software employed to collect data from monkeys in this study was, unless otherwise specified, identical to that described by Jensen et al. (2013a).

**Human Participants**

41 students enrolled at Columbia University participated in the study to fulfil an introductory psychology class requirement. Participants were given minimal verbal instruction. They were told only (1) that they were to use a mouse to click on images, (2) that feedback would consist of green check marks (indicating a correct response) or red crosses (indicating an incorrect responses), and (3) that they should try to get as many correct responses as possible. No mention of either “serial” or “categorical” cognition was made until participants were debriefed.

Participants first completed 120 trials of the Category Chain task using the four photographic categories (using the order Birds → Cats → Flowers → Humans). They then completed 200 trials in which stimuli were derived from works by four painters (Dalí → Gérôme → Monet → Van Gogh). Unlike the monkeys, participants were given no prior training or instruction regarding the task structure of the category memberships. They had to learn how to categorize the stimuli at the same time as they learned the task demands. Consequently, participants generally began by responding at chance and gradually learned what the correct responses were over successive trials.

**CALCULATION**

In traditional models of choice, each trial consists of a single choice and choice is modeled as the probability of selecting a particular stimulus. The analysis of SimChain (and, by extension, Category Chain), is complicated, however, by the varying number of individual responses during each trial. For example, when presented with a 4 item list, a subject might make one, two, three, or four responses, depending on whether any erroneous responses were made in the sequence. Consequently, the analysis of Category Chain performance requires the simultaneous estimation of a different conditional probability for each response in the sequence.

Let \( p_1 \) correspond to the probability of a correct first response, \( p_2 \) to the probability of a correct second response, and so forth. The probability of reward in any single trial of the Category Chain depended on four probabilities:

\[
p(\text{reward}) = p_1 \cdot p_2 \cdot p_3 \cdot p_4
\]

For any intermediate degree of progress, the chain of probabilities is cut off following an error (whose probability is \((1 - p_i)\) for choice \(i\)). For example, if only the first two responses are correct and the third response is an error, then:

\[
p(\text{progress} = 2) = p_1 \cdot p_2 \cdot (1 - p_3)
\]

In order to characterize performance in the Category Chain task, we developed a formal model of subjects’ individual responses. This chain of conditional probabilities is fully depicted in Figure 2A.
Figure 2. **Formal model for describing simultaneous chaining behaviour.** (A). The decision tree that govern an animal’s “progress” through the list in a single trial. Each choice made by the subject has a probability of being correct $p_1, p_2, \ldots$, such that the probability of completing a 4-item list (and thus earning a reward) is $(p_1 \cdot p_2 \cdot p_3 \cdot p_4)$. Since specifying each permits the likelihood to be calculated, posterior distributions for these probabilities can be estimated. (B). The asymmetric bounded logistic function, used to model each choice probability $p_i$ in humans (who lacked prior experience, and so had to learn the categories while doing the task). This function is defined in terms of peak learning rate (governed by $m$), a slope (governed by $s$), a floor term denoting starting performance (governed by $f$), and a twist parameter that influenced the asymmetry in learning speeds early vs. late during learning (governed by $v$). (C). Examples of the conditional probabilities $p_i$ for each of the four choices. Note that $p_3$ begins higher than $p_1$ because, should a subject get to the third choice, only two items will remain, making it a 50/50 chance. $p_4$ is modeled as a constant value near 1.0. (D). Joint probabilities associated with reaching different choice points, using the four probability function in panel C. (E). Average expected progress in the list, computed by taking the sums of the joint probabilities in panel D. Thus, given a function for each choice probability $p_i$, one can also model the progress made by participants.

Performance can be precisely described if each of these probability parameters can be estimated. Because the monkeys received extensive training prior to test, their performance had already achieved a stable ceiling. Consequently, the probability of a correct choice to stimulus was estimated using the logit link:

$$p_i = \frac{1}{1 + \exp(-m_i)}$$

(1)

The human data presented a more complicated analysis problem, however, because they had no prior training on this task. Thus, each probability was expected to begin at chance, then grow toward 1.0 over time. To characterize learning, we used a variant of the generalized logistic function (Richards, 1959), which we call the asymmetric bounded logistic:

$$p_i(t) = f + \frac{1 - f}{1 + \exp(-[s_i \cdot t + m_i]^v_i)}$$

(2)

Here, $p_i(t)$ refers to the probability of a correct response $i$ to stimulus at time $t$. The function then accepts four parameters. The slope ($s$) influences the speed at which learning unfolds, and the peak ($m$) influences when learning begins to differ from chance. The level of chance performance for each stimulus is in turn
governed by a floor parameter ($f$). Finally, a twist parameter ($v$) is included because past empirical work on SimChain suggests that once learning begins, it improves very rapidly at the outset, with diminishing returns as performance reaches asymptote (Jensen et al., 2013). A depiction of the asymmetric bounded logistic is presented in Figure 2B, along with a description of the contributions of each parameter.

Figure 2C shows hypothetical learning curves for each response over the course of 120 trials. The logistic functions for $p_1$, $p_2$, and $p_3$ all have the same parameters for slope ($s = 0.1$) and twist ($v = e^6$), but differ in their values for the floor $f$ and the peak $m$. The floors differ because, when a participant uses process-of-elimination search, chance performance given $n$ stimuli is $\frac{1}{n}$ for the first item (because there is a 1 in $n$ chance of choosing the correct alternative), $\frac{1}{n-1}$ for the second item, and so forth. The intercepts differ because progress cannot be made on discovering the identity of Item 2 until some learning regarding Item 1 has begun. Finally, the last probability $p_4$ is a constant close to 1.0, fit using Equation 1, because subjects almost never make mistakes on the final choice, again thanks to process-of-elimination search.

The need for asymmetry in the learning rate (with a long period at some floor value, followed by a sharp acceleration and then gradual diminishing returns) stems from SimChain’s process of sequential discovery. Prior to discovering the identity of the first stimulus, subjects cannot gain information about the remaining stimuli. It is only after the first stimulus was selected with some consistency that any information about later list items can be discovered. This is why the conditional probability for $p_3$ remains flat for 50 trials in Figure 2C. It is only after the first two stimuli have been acquired that the subject has an opportunity to perform above chance with respect to the third stimulus.

Figure 2D depicts the joint probability of the same participant making progress through the list. The red line is the same as in Figure 2C, depicting how often the first choice is correctly made. The blue line, however depicts how often the participant got at least two choices correct (determined by multiplying $p_1$ and $p_2$), the green line at least three correct, and so forth. Figure 2E depicts the mean expected progress in the list, and is determined by taking the sum of the joint probabilities in Figure 2D. In this case, the participant begins at chance levels ($progress = 0.4167$), but is earning a reward for almost every trial by trial 120.

Parameter estimates were obtained using the Stan language (Carpenter et al., 2015). The analysis script is included as supplementary information.

RESULTS

We analyzed the behavior of monkeys and humans performing the Category Chain task with respect to four photographic categories and four painting categories. In particular, our analysis focused on making estimates of the conditional probabilities of the responses to each stimulus, and their corresponding reaction times.

Monkeys

Figure 3A depicts the conditional probabilities for the three monkeys who learned to classify photographic stimuli. Figure 3B depicts their overall probability of obtaining a reward and Figure 3C depicts their mean progress in the list. The violin plots correspond to the posterior distribution of each estimate. Despite the presentation of different stimuli during every trial, accuracy was reliably high. Subjects responded correctly to at least 75% of the first-position stimuli (as compared to chance accuracy of 25%); at least 83% of the second-position stimuli (chance accuracy = 33%); and at least 85% of third-position stimuli (chance accuracy = 50%). Jointly across trials, this led to rewards being earned on roughly half the trials.

Figure 4 depicts the log reaction times of the photographic monkeys, displaying both the overall
Figure 3. Monkey Category Chain performance for photographic stimuli. In all cases, violin plots represent the posterior density for the parameter estimate. (Left) Conditional response probabilities for each subject with respect to each list item. (Middle) Overall probability that any given trial will result in reward delivery (equal to the product of the conditional probabilities) for each subject. (Right) Mean progress in the list on each trial for each subject.

Figure 4. Monkey reaction times to photographic stimuli on a log scale. Violin plots show the distribution of log reaction times (in gray) and the credible interval for the mean reaction time (in white) for each monkey.

variability of reaction times (gray violins) and the credible interval for the mean (white violins). Subjects consistently made each response during a trial faster than the last, as would be expected using a process-of-elimination strategy.

Figure 5A depicts the conditional probabilities, Figure 5B depicts overall probability, and Figure 5C depicts mean performance for the four monkeys who classified painterly stimuli. Although all four animals displayed conditional probabilities above chance, the overall accuracy was consistently lower than that observed for photographic classification. As a result, only about one trial in three ended with reward.

Figure 6 tells a similar story about the log reaction times: Subjects responded more rapidly to later list items than to earlier ones, but tended to respond more slowly overall. One animal in particular, Prospero, tended to respond at only about 1/3 as rapidly as did the photographic group (cf. Figure 4). This suggests that the painterly stimuli were more difficult to classify, relative to the comparatively automatic classification of the photographic stimuli.

Overall, these results resemble patterns of learning observed in classical SimChains presenting fixed arbitrary photographs (Terrace, 2005; Jensen et al., 2013a), and in SimChains for which the list order was determined by some psychophysical dimension (Avdagic et al., 2014). This demonstrates both that subjects were able to classify the categorical stimuli, and that the serial learning of categories was consistent with
**Figure 5.** Monkey Category Chain performance for painting stimuli. In all cases, violin plots represent the posterior density for the parameter estimate. (Left) Conditional response probabilities for each subject with respect to each list item. (Middle) Overall probability that any given trial will result in reward delivery (equal to the product of the conditional probabilities) for each subject. (Right) Mean progress in the list on each trial for each subject.

**Figure 6.** Monkey reaction times to painting stimuli on a log scale. Violin plots show the distribution of log reaction times (in gray) and the credible interval for the mean reaction time (in white) for each monkey.

other previously reported forms of serial learning. What distinguishes the current task from previous studies of serial learning (e.g. using classical SimChain) is that every stimulus changed on every trial. This meant that subjects had to classify the stimuli before they could determine their serial order. Additional details, including distributions of estimated parameters, are provided in the supplementary materials.

**Humans**

Figure 7 (top) depicts the best-fitting learning curves, using Equation 2, of individual human participants during the photographic phase (semitransparent lines), as well as their group mean (solid black line). Additionally, the empirical trial means are plotted as white points. These learning curves indicate that most participants reached near-perfect performance (classifying all four stimuli correctly on nearly every trial) within the first 40 trials of 120. The steep learning curves displayed by most participants suggests that once the task demands were understood, discovering the order unfolded very rapidly.

Figure 7 (bottom) confirms this suspicion, plotting the population density function for the conditional probabilities (gray violins) as well as the posterior distributions for mean performance. Performance at the start of training (t = 0) looks just as one would expect for chance performance, but by the session’s midpoint (t = 60), a bimodal distribution had emerged of predominantly successful participants with a handful still responding at chance levels. By the end of the session, the bimodal data continue to drag the
Figure 7. Human performance estimate for photographic stimuli. (Top). Estimated time course of Category Chain performance for individual participants (gray curves) and the overall mean (black curve). White points correspond to the empirical mean of performance across subjects on each trial. Chance is indicated by the dashed line. (Bottom). Estimated conditional response probability for each list item at session onset ($t = 0$), midway through the session ($t = 60$), and after the last trial ($t = 120$). Violin plots show the distribution of individual probability estimates (in gray) and the credible interval for the mean probability across participants (in white). Note that mean probabilities are usually much lower than the population mode (which is near 1.0), due to a subset of participants who remain at chance throughout training.

The overall mean down somewhat, but the mode is so close to 1.0, as the participants who are making close to zero errors dominate the distribution. The rapid acquisition by humans is consistent with an account whereby they identify the categories using familiar linguistic labels.

Figure 8 tells a similar story with regards to the reaction times. Figure 8 (left) depicts the empirical means of the log reaction times as points for first responses (white points), as well as the second (light gray), third (dark gray), and fourth (black) responses. With the exception of very early trials, these were approximately linear, so a hierarchical linear regression was performed (black lines). These results suggest that, at the outset of training, responses displayed the characteristic acceleration seen in monkeys, in which each response was made more quickly than the last (compare to Figures 4 and 6). This pattern is consistent with a visual search strategy, working by process of elimination one response at a time. However, by the end of the session, a different pattern had emerged: A long interval for the first responses, followed by rapid selection of the second through fourth items. This pattern is more characteristic of a plan-then-execute approach. This pattern is also evident in Figure 8 (right), which shows the posterior population estimates (gray violins) and credible intervals for the line of best fit (white violins) at both the
Figure 8. Human reaction times to photographic stimuli on a log scale. (Left). Regression model of estimate first touch (white), second touch (light gray), third touch (dark gray), and final touch (black). Right). Estimated reaction time at session onset ($t = 0$) and after the last trial ($t = 120$) for touches to each list item. Violin plots show the distribution of log reaction times (in gray) and the credible interval for the mean reaction time (in white).

start ($t = 0$) and the end ($t = 120$) of training. Estimates showed a consistent downward trend at the start of training. Reaction times at the end of training were very similar after the first response.

Figure 9 (top) depicts the performance of the same participants, in terms of Equation 2, to the painting phase of the experiment. Unlike Figure 7, participants did not show rapid acquisition, instead improving only gradually as the session unfolded. Similarly, Figure 9 (bottom) shows that the population distributions of conditional probabilities do not converge rapidly toward ceiling as they did in the photographic condition. Although participants did tend to improve to varying degrees (including a few who were able to classify all four stimuli), many improved so slowly as to be indistinguishable from chance.

Figure 10 (left) depicts reaction times over the course of the painting phase. A plan-then-execute approach is no longer evident, with reaction times for the second and third items remaining consistently long and similar to one another. The third response (in which participants must distinguish Monet from van Gogh) appeared especially difficult, emerging as slower than the second response by the end of the session.

Overall, these results provide a dramatic contrast. In the photographic phase, despite receiving no verbal instruction, participants were able (1) to realize that stimuli belonged to regular categories, (2) to learn that rewards depended on selecting categories in a particular order, and (3) to classify stimuli consistently. Within 60 trials, most previously-naive participants were classifying stimuli with near-perfect accuracy (Figure 7), whereas highly trained monkeys performing the same task and evaluating the same stimuli only managed to earn rewards on about 50% of trials (Figure 3). In the painting phase, however, humans demonstrated much greater difficulty identifying the categories and thus classifying the stimuli, despite being familiar with the task. By the end of 200 trials, humans (Figures 9) still resembled monkeys (Figures 5), classifying the first three items with approximately 70% accuracy.

**DISCUSSION**

Our results compare performance on two primate species, macaques and humans, across two types of visual categories: naturalistic photographs and famous artists’ paintings. Category exemplars were selected from large and highly disparate sets of images. Accuracy during the painting phase was lower for both monkeys and humans, indicating that the paintings were more difficult to categorize. Nevertheless, both species were able to successfully learn and simultaneously classify stimuli from both sets of categories.
Figure 9. Human performance estimate for photographic stimuli. (Top). Estimated time course of Category Chain performance for individual participants (gray curves) and the overall mean (black curve). White points correspond to the empirical mean of performance across subjects on each trial. Chance is indicated by the dashed line. (Bottom). Estimated conditional response probability for each list item at session onset ($t = 0$), midway through the session ($t = 60$), and after the last trial ($t = 120$). Violin plots show the distribution of individual probability estimates (in gray) and the credible interval for the mean probability across participants (in white). Note that mean probabilities are usually much lower than the population mode (which is near 1.0), due to a subset of participants who remain at chance throughout training.

This categorization is noteworthy for two reasons. Firstly, our task (the Category Chain) was unusually demanding, requiring that subjects simultaneously classify four stimuli. Secondly, both the photographic stimuli and the painting stimuli were large and diverse. The paintings, in particular, could not be categorized by attending to a handful of discrete features. These two factors (task complexity and stimulus set complexity) jointly provide compelling evidence that subjects processed stimuli as gestalt percepts rather than resorting to a simple feature-based strategy. This permitted subjects to classify stimuli even when category membership depended on high-level stimulus properties rather than low-level image statistics or specific feature discriminations.

The photographs are representative of ecological stimuli; humans are familiar with these categories, and we possess unambiguous linguistic scaffolds for each. Although captive monkeys have little exposure to wildlife and do not possess language, primates nevertheless evolved in the presence of birds, flowers, cats, and other primates, and past experiments have demonstrated a greater aptitude at classifying naturalistic photographs than those of man-made objects (Jitsumori and Matsuzawa, 1991).

While our monkeys performed well above chance with photographic stimuli, they made consistent,
systematic errors, even after thousands of trials of training. By comparison, most human participants were able to perform the task with close to zero errors after less than 100 trials. Many studies have reported that monkeys tend to make errors, even when the discriminations are simple and the animals experienced. One hypothesis for this persistent error rate is that while both monkeys and humans face a speed-accuracy tradeoff, monkeys tend to favor speed while humans tend to favor accuracy (Richard and Thorpe, 1998; Heitz and Schall, 2012). This does not seem like a satisfactory account of the present data, however, because monkeys and humans had similar reaction times by the end of training (Figure 4 vs. Figure 8, Figure 6 vs. Figure 10). Another possible reason for differences in response accuracy may stem from species differences in executive control. Monkeys tend in general to behave under the influence of stimuli, whereas humans tend to behave under the influence of more abstract task demands (Stoet and Snyder, 2003; Brosnan et al., 2013). Because the photographic stimulus categories were familiar to human participants, recognition appeared fairly automatic and was unperturbed by changes in the stimuli from trial to trial. If the monkeys, however, made more deliberate assessments of the stimuli on every trial, it would help explain persistent errors.

The paintings are not representative of any stimuli that are frequently encountered in the wild and were more difficult to classify. When classifying the paintings, monkeys were less accurate and their systematic errors were more frequent. Humans also found the paintings phase more difficult, making more errors and learning more slowly (Figure 7 vs. Figure 9). However, monkey reaction times to paintings did not markedly differ from reaction times to photographs. This suggests that monkeys used a similar classification strategy with the paintings stimuli as with the photographic stimuli.

Unlike the monkeys, humans displayed two patterns of response. The photographic stimuli elicited a “single slow response, followed by three quick responses” pattern (consistent with planning an entire sequence before executing the plan), whereas the painting stimuli forced participants to pause at each decision. In this respect, humans differed from monkeys when classifying photographs, but resembled them when classifying paintings. The differing patterns of reaction times and the very slow improvement in performance suggest that human participants had no existing linguistic representations to link to these percepts. Participants may have seen full images of some of the paintings before, but by zooming in on the details and brushwork of the paintings, we attenuated whatever benefits would have come with familiarity with the artists’ works. When faced with sufficiently abstracted categories, humans seem to rely on a

---

**Figure 10. Human reaction times to painting stimuli on a log scale.** (Left). Regression model of estimate first touch (white), second touch (light gray), third touch (dark gray), and final touch (black). (Right). Estimated reaction time at session onset ($t = 0$) and after the last trial ($t = 120$) for touches to each list item. Violin plots show the distribution of log reaction times (in gray) and the credible interval for the mean reaction time (in white).
One of the chief difficulties in determining how features are used by classifiers is that it is unclear what constitutes a “feature” in objective terms. For example, the plausibility of the hypothesis that photographs are classified based on discrete features (such as eyes and beaks) depends on the ease with which we are able to identify such features. Traditionally, features were either defined in terms of “bottom-up” or “top-down” information (Wolfe et al., 2003). A bottom-up classifier identifies a feature through the assemblage of low-level sensory features (such as on-center receptive fields and edges). Such systems are thus strictly driven by primary sensory information, without any influence of prior expectation. Top-down classifiers, on the other hand, are heavily influenced by prior knowledge, and act on sensory representations that have already undergone several rounds of synthesis and integration. It is very difficult to provide direct evidence of top-down processing in animals (Tomita et al., 1999), and as a result the study of animal classifiers has focused almost exclusively on bottom-up processing (Vogels, 1999; Marsh and MacDonald, 2008; Srihasam et al., 2014). Although bottom-up processing is clearly involved in how the brain classifies images, it is far from clear that it is sufficient to explain the abilities of non-human animals.

In order to compare our empirical data to a purely bottom-up approach, we simulated task performance using the “bag-of-features” image classifier (O’Hara and Draper, 2011). This sophisticated algorithm makes discriminations only on the patterns of low-level statistical regularities. Insofar as subjects outperformed the algorithms, we take this as evidence that subjects’ strategy was more than merely bottom-up, instead involving at least some hierarchical processing.

Figure 11 shows the mean probability of a correct response for monkeys (in white), humans at the end of the discrimination task, and the classifier (in gray). The correct response probabilities for monkeys (white points) and humans (black points) for each of conditional probabilities $p_1$ to $p_4$ include error bars for the 95% credible interval of the estimate. Accuracy is also included for the bag-of-features classification algorithm (gray). (Left). Performance in classifying stimuli from the four photographic categories. Humans systematically outperformed monkeys, who in turn outperformed the algorithm. (Right). Performance in classifying stimuli from the four painting categories. Neither humans and monkeys were consistently superior, although both outperformed the algorithm on the difficult initial discrimination.

Figure 11. Correct response probabilities for monkeys, humans, and a computer vision algorithm. Estimated response probabilities for monkeys (white points) and humans (black points) for each of conditional probabilities $p_1$ to $p_4$ include error bars for the 95% credible interval of the estimate. Accuracy is also included for the bag-of-features classification algorithm (gray). (Left). Performance in classifying stimuli from the four photographic categories. Humans systematically outperformed monkeys, who in turn outperformed the algorithm. (Right). Performance in classifying stimuli from the four painting categories. Neither humans and monkeys were consistently superior, although both outperformed the algorithm on the difficult initial discrimination.
of one session (in black), and the bag-of-feature algorithm (in gray) when presented with both photographs and paintings. The algorithm probabilities assume a process-of-elimination search with no backwards errors. Of particular importance was the probability of selecting the first item correctly (where chance performance would be 0.25), since this was the case in which the target had to be distinguished from three distractors. For the photographic stimuli, humans clearly outperformed monkeys, but monkeys also outperformed the algorithm. The paintings, however, yielded lower performance for all cases, confirming the intuition that the paintings ought to be more difficult to differentiate. In addition, although humans and monkeys outperformed the bag-of-features algorithm on the first painting discrimination, neither species clearly outperformed the other.

The similarity between monkey and human performance in the case of the paintings distinguishes our stimuli from sets of man-made stimuli used in other primate studies (Sigala, 2009; Fize et al., 2011), which incorporated consistent features, like wheels or windows. Although paintings have been used as discriminative stimuli in animals studies by Watanabe and colleagues (Watanabe, 2001; Ikkatai and Watanabe, 2011; Watanabe, 2013), ours is the first study to use painting stimuli with non-human primates. This study also presented subjects with a much more difficult task than previous studies. Not only did subjects have to classify stimuli from four different categories simultaneously, but did so using only small segments of the original paintings rather than the full canvas. The difficulty of the task and the ambiguity of the stimuli both helped to ensure that categorization performance could not be ascribed to simple associative mechanisms. Most studies of animal categorization require that subjects discriminate between only two categories, and such tasks can be solved using shortcuts that would fail if more than two categories were presented simultaneously (Jensen and Altschul, 2015).

The ability to categorize percepts has been found in multiple species of primate (Schrier et al., 1984; Vonk and MacDonald, 2002; Marsh and MacDonald, 2008), as well as pigeons (Bhatt et al., 1988; Roberts and Mazmanian, 1996; Watanabe, 2001), so it should not be surprising that non-human primates can flexibly classify stimuli according to abstract stimulus properties. Additionally, both humans and macaques have previously been shown to attend to the overall family resemblance of sets of stimuli, even in cases when doing so impairs performance because the task requires selecting on the basis of only a single feature (Couchman et al., 2010).

Our study has several limitations, which open numerous avenues for future work. We explicitly provided as little information as possible to our human participants in advance of the experimental task. For example, if human participants were trained in art history, then having a formal framework should permit linguistic classification, with a corresponding improvement in performance. None of our participants were art majors, so it might also be revealing to compare trained artists or art historians to amateurs, to see if formal training would yield performance more in line with the rapid learning displayed with the photographic stimuli.

Identifying the specific operations of the classifiers that monkeys use remains a formidable problem. Categorization is a rich sub-field of comparative cognition, but we believe that future investigations will benefit from more demanding test procedures, including larger stimulus sets, more categories, and harder tasks. Additionally, it is important to demonstrate these results in a variety of contexts, i.e., in multiple species and with different experimental tasks. The Simultaneous Chaining paradigm is known to map onto a common serial representation (Jensen et al., 2013a), making the task both generalizable and sufficiently difficult. Nevertheless, the abstract classification abilities demonstrated by the monkeys in this study ought to be replicated with other cognitive testing paradigms.

Understanding complex learning in animals has always been difficult because animals cannot learn human languages. Our results suggest that stimuli can be integrated into percepts and can in turn be classified without the benefit of language. An animal’s lack of human language is most often seen as an
obstacle, but when manipulated appropriately, non-linguistics animals, and comparative psychology in general, can teach us much about the structure of cognition.

ACKNOWLEDGMENTS

The authors wish to thank Adebayo Adesomo, Farhana Begum, Shangshang Chen, Daniel Deihle, and Rachelle Meyer for their assistance with the monkeys, and Vincent Ferrera for feedback on an earlier version of this manuscript. This work was supported by US National Institute of Health, grant 5R01MH081153-06 awarded to H.T.

AUTHOR CONTRIBUTIONS

D.A., G.J., and H.S. conceived the experiments. D.A. and G.J. wrote the task software, acquired data, and performed analyses. D.A., G.J., and H.S. wrote the paper.

REFERENCES

Avdagic, E., Jensen, G., Altschul, D., and Terrace, H. S. (2014). Rapid cognitive flexibility of rhesus macaques performing psychophysical task-switching. Animal Cognition, 17:619–631.
Bhatt, R. S., Wasserman, E. A., Reynolds, W. F., and Knauss, K. S. (1988). Conceptual behavior in pigeons: Categorization of both familiar and novel examples from four classes of natural and artificial stimuli. Journal of Experimental Psychology: Animal Behavior Processes, 14:219–234.
Brosnan, S. F., Beran, M. J., Parrish, A. E., Price, S. A., and Wilson, B. J. (2013). Comparative approaches to studying strategy: Towards an evolutionary account of primate decision making. Evolutionary Psychology, 11:606–627.
Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2015). Stan: A probabilistic programming language. Journal of Statistical Software.
Chater, N. and Heyes, C. (1994). Animal concepts: Content and discontent. Mind and Language, 9:209–246.
Close, J., Hahn, U., Hodgetts, C. J., and Pothos, E. M. (2010). Rules and similarity in adult concept learning. In Mareschal, D., Quinn, P. C., and Lea, S. E. G., editors, The Making of Human Concepts, pages 29–51. Oxford University Press, Oxford, UK.
Couchman, J. J., Coutinho, M. V. C., and Smith, J. D. (2010). Rules and resemblance: their changing balance in the category learning of humans (homo sapiens) and monkeys (macaca mulatta). Journal of Experimental Psychology: Animal Behavior Processes, 36:172–183.
Escher, M. C. (1989). Escher on Escher: Exploring the Infinite. Harry N. Adams, New York, NY.
Fize, D., Caichoix, M., and Fabre-Thorpe, M. (2011). Humans and monkeys share visual representations. Proceedings of the National Academy of Sciences of the USA, 108:7635–7640.
Flemming, T. M., Thompson, R. K., and Fagot, J. (2013). Baboons, like humans, solve analogy by categorical abstraction of relations. Animal Cognition, 16:519–524.
Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science, 291:312–316.
Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., and Gore, J. C. (1997). Levels of categorization in visual recognition studied using magnetic resonance imaging. Current Biology, 7:645–651.
Goldstone, R. L. and Kersten, A. W. (2003). Concepts and categorization. In Healy, A. F., Proctor, R. W.,
and Weiner, I. B., editors, *Handbook of Psychology, Vol. 4: Experimental Psychology*, pages 599–621. Wiley, New York, NY.

Griffin, G., Holub, A. D., and Perona, P. (2007). The Caltech 256. Technical Report CNS-TR-2007-001, California Institute of Technology.

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

Hampton, J. A. (2003). Concepts in human adults. In Mareschal, D., Quinn, P. C., and Lea, S. E. G., editors, *The Making of Human Concepts*, pages 296–314. Oxford University Press, Oxford, UK.

Heitz, R. P. and Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, 76:616–628.

Herrnstein, R. J. (1990). Level of stimulus control: A functional approach. *Cognition*, 37:133–166.

Herrnstein, R. J., Loveland, D. H., and Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 2:285–302.

Huber, L. (2000). Generic perception: Open-ended categorization of natural classes. In Fagot, J., editor, *Picture Perception in Animals*, pages 219–261. Psychology Press, Hove, East Sussex.

Ikkatai, Y. and Watanabe, S. (2011). Discriminative and reinforcing properties of paintings in java sparrows (padda oryzivora). *Animal Cognition*, 14:227–234.

Jensen, G. and Altschul, D. (2015). Two perils of binary categorization: Why the study of concepts can’t afford true/false testing. *Frontiers in Psychology*, 6:Article 168.

Jensen, G., Altschul, D., Danly, E., and Terrace, H. S. (2013a). Transfer of a serial representation between two distinct tasks by rhesus macaques. *PLOS ONE*, 8:e70285.

Jensen, G., Ward, R. D., and Balsam, P. D. (2013b). Information: Theory, brain, and behavior. *Journal of the Experimental Analysis of Behavior*, 100:408–431.

Jitsumori, M. and Delius, J. D. (2001). Object recognition and object categorization in animals. In Matsuzawa, T., editor, *Primate Origins of Human Cognition and Behavior*, pages 269–293. Springer, New York, NY.

Jitsumori, M. and Matsuzawa, T. (1991). Picture perception in monkeys and pigeons: transfer of right-side-up versus upside-down discrimination of photographic objects across conceptual categories. *Primates*, 32:473–482.

Katz, J. S., Wright, A. A., and Bodily, K. D. (2007). Issues in the comparative cognition of abstract-concept learning. *Comparative Cognition and Behavior Reviews*, 2:79–92.

Lea, S. E. G. (1984). In what sense do pigeons learn concepts? In Roitblat, H. L., Bever, T. G., and Terrace, H. S., editors, *Animal Cognition*, pages 263–276. Erlbaum, Hillsdale, NJ.

Lee, S. M., Xin, J. H., and Westland, S. (2005). Evaluation of image similarity by histogram intersection. *Color Research and Application*, 30:265–274.

Leopold, D. A. and Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature*, 379:549–553.

Marsh, H. L. and MacDonald, S. E. (2008). The use of perceptual features in categorization by orangutans (Pongo abelli). *Animal Cognition*, 11:569–585.

Matsukawa, A., Inoue, S., and Jitsumori, M. (2001). Pigeon’s recognition of cartoons: effects of fragmentation, scrambling, and deletion of elements. *Behavioural Processes*, 65:25–34.

Miller, E. K., Nieder, A., Freedman, D. J., and Wallis, J. D. (2003). Neural correlates of categories and concepts. *Current Opinion in Neurobiology*, 13:198–203.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*.

O’Hara, S. and Draper, B. A. (2011). Introduction to the bag of features paradigm for image classification and retrieval. *arXiv Preprints*, page 1101.3354.
Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. In IEEE Conference on Computer Vision and Pattern Recognition.

Pezaris, J. S. and Reid, R. C. (2007). Demonstration of artificial visual percepts generated through thalamic microstimulation. Proceedings of the National Academy of Sciences USA, 104:7670–7675.

Reddy, L., Tsuchiya, N., and Serre, T. (2010). Reading the mind’s eye: Decoding category information during mental imagery. NeuroImage, 50:818–825.

Richard, M. F.-T. G. and Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus macaques. NeuroReport, 9:303–308.

Roberts, W. A. (1996). Stimulus generalization and hierarchical structure in categorization by animals. Advances in Psychology, 117:35–54.

Roberts, W. A. and Mazmanian, D. S. (1996). Concept learning at different levels of abstraction by pigeons, monkeys, and people. Journal of Experimental Psychology: Animal Behavior Processes, 14:247–260.

Schrier, A. M., Angarella, R., and Povar, M. L. (1984). Studies of concept formation by stumptailed monkeys: Concepts humans, monkeys, and letter A. Journal of Experimental Psychology: Animal Behavior Processes, 10:564–584.

Schrier, A. M. and 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, 13:136–143.

Sigala, N. (2009). Natural images: A lingua franca for primates? Open Neuroscience Journal, 3:48–51.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J., and Ashby, F. G. (2010). Implicit and explicit category learning by macaques (macaca mulatta) and humans (homo sapiens). Journal of Experimental Psychology: Animal Behavior Processes, 36:54–65.

Sovranò, V. A., Albertazzi, L., and Salva, O. R. (2010). The ebbinghaus illusion in a fish (xenotoca eiseni). Animal Cognition, 18:533–542.

Srihasam, K., Vincent, J. L., and Livingstone, M. S. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. Nature Neuroscience, 17:1776–1783.

Stoet, G. and Snyder, L. H. (2003). Executive control and task-switching in monkeys. Neuropsychologia, 41:1357–1364.

Terrace, H. S. (1984). Simultaneous chaining: The problem it poses for traditional chaining theory. In Commons, M. L., Herrnstein, R. J., and Wagner, A. R., editors, Quantitative Analyses of Behavior: Discrimination Processes, pages 115–138. Ballinger, Cambridge, MA.

Terrace, H. S. (2005). The simultaneous chain: A new approach to serial learning. TRENDS in Cognitive Sciences, 9:202–210.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., and Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature, 401:699–703.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. part 2: single-cell study. European Journal of Neuroscience, 11:1239–1255.

Vonk, J. (2013). Matching based on biological categories in orangutans (Pongo abelii) and a gorilla (Gorilla gorilla gorilla). PeerJ, 1:e158.

Vonk, J. and 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, 78:315–332.

Wasserman, E. A., Brooks, D. I., and McMurray, B. (2015). Pigeons acquire multiple categories in parallel via associative learning: A parallel to human word learning? Cognition, 136:99–122.

Wasserman, E. A., Fagot, J., and Young, M. E. (2001). Same-different conceptualization by baboons (Papio papio): The role of entropy. Journal of Comparative Psychology, 115:42–52.
Watanabe, S. (2001). Van gogh, chagall, and pigeons: Picture discrimination in pigeons and humans. *Animal Cognition*, 4:147–151.

Watanabe, S. (2013). Preference for and discrimination of paintings by mice. *PLOS ONE*, 8:e65335.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2010). Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Wolfe, J. M., Butcher, S. J., Lee, C., and Hyle, M. (2003). Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology: Human Perception and Performance*, 29:483–502.

Wright, A. A. and Lickteig, M. T. (2010). What is learned when concept learning fails? a theory of restricted-domain relational learning. *Learning & Motivation*, 41:273–286.

Young, M. E. and Wasserman, E. A. (2002). The pigeon’s discrimination of visual entropy: A logarithmic function. *Animal Learning and Behavior*, 30:306–314.

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., and Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition & Behavior Reviews*, 3:13–45.
APPENDIX A: STIMULUS ANALYSIS

Although comparative studies of primate cognition very often use photographic stimuli, systematic analyses of the stimuli are rarely undertaken. This is unfortunate, because skepticism about surprising results often relies on speculation about stimulus characteristics that might have been used as discriminative cues. A subject might, for example, identify pictures of birds solely on the basis of blue backdrops (i.e. the sky). Analysis can determine whether bird stimuli contain disproportionate amounts of blue.

Our analysis of stimulus images focuses on low-level features (e.g. color, image entropy). This approach has the advantage of being automatic and replicable. The overarching question that these analyses seek to inform is this: “To what extent are low-level properties sufficient to categorize stimuli correctly?” More diverse stimuli in each grouping means more difficulty specifying criteria for category inclusion. At the same time, the more each of the categories resembles the others overall, the more difficult it is to specify criteria for category exclusion.

The stimuli used in this study were selected in a fashion that differs from the conventions used in typical psychophysical experiments. Rather than select stimuli according to strict inclusion criteria, or modifying images before use (e.g., turning them grayscale or giving them uniform spectra), we included images solely on the basis of the question, “Is this a picture of X?” For example, our photographs of people included both extreme close-ups of faces and wide-angle views of crowds. We also included both color and black-and-white images.

The photographic stimulus sets were birds (3111 images), cats (1114 images), flowers (3033 images), and people (2853 images). The painting stimulus sets were derived from works by Dalí (400 images), Gérôme (406 images), Monet (400 images), and van Gogh (400 images).

Figures 12 depict a representative set of exemplars for each of the painting categories. For reasons associated with image and likeness rights, exemplars from the photographic categories are not included. However, representative images may be obtained from the Caltech-UCSD Birds 200 Dataset (Welinder et al., 2010) for the ‘birds’ category, from the Oxford-IIIT Pet Dataset (Parkhi et al., 2012) for the ‘cats’ category, from the Oxford 102 Category Flower Dataset (Nilsback and Zisserman, 2008) for the ‘flowers’ category, and from the Caltech 256 Dataset (Griffin et al., 2007) for the ‘people’ category. Subjects with prior experience using the SimChain paradigm were not previously exposed to any of the images in each of the categorical image banks.

Pixel Entropy

Prior research has shown that primates possess the ability to discriminate stimuli based on visual entropy Flemming et al. (2013); Wasserman et al. (2001), an ability also demonstrated in pigeons Young and Wasserman (2002). Because the entropy estimation can be done mechanically by simple systems, doing so falls considerably short of the criteria for a ”categorical representation.” Consequently, an analysis of pixel entropies gives an idea of whether the sets of stimuli differ sufficiently to be discriminated on that basis.

Here, pixel entropy is taken to be the Shannon entropy Jensen et al. (2013b), computed over all possible combinations of red, blue, and green intensities:

\[
H = \sum_{r=0}^{255} \sum_{g=0}^{255} \sum_{b=0}^{255} p(r, g, b) \log_2 (p(r, g, b))
\]  

The maximum possible entropy \(H\) that a bitmap image could possibly display is 24, provided each of the \(256 \times 256 \times 256\) pixel values appears equally. However, such an entropy would require a \(4096 \times 4096\)
Figure 12. Exemplars of the stimuli drawn from the works of four painters, used with monkeys and humans.
pixel image, much larger than our stimuli. Because our stimuli were only \(200 \times 200\) pixels in size, the highest possible entropy that a color stimulus could possess was 15.29 bits. Grayscale images had a maximum entropy of 8 bits.

Figures 13 and 14 show kernel density estimates of the distributions of pixel entropies displayed in the photographic and painterly categories, respectively, as well as each distribution’s quartiles. In general, stimuli tended to show high entropies of between 12 and 14 bits, such that a 13-bit image could easily belong to any of the categories. However, the photographic stimuli show clear distributional differences. For example, many more of the images of birds have entropies below twelve than the other stimuli, while the images of flowers routinely have higher entropies than the other stimuli.

The painting stimuli tend to have higher entropies overall than the photographic stimuli. Here, too, however, there are notable similarities. Dalí and Gérôme both resemble one another closely, as do Monet and van Gogh, but these two clusters appear distinct from one another. Importantly, however, because stimuli in each of these pairings are distributed so similarly, it would be very difficult for subjects to distinguish each group precisely on the basis of pixel entropy alone.

We do not rule out the possibility that pixel entropy facilitated identification in some fashion. This analysis is merely intended to demonstrate that pixel entropy alone would not have been sufficient to precisely classify each stimulus.

**HSV Histograms**

Another method by which images can be compared is on their HSV distributions. Just as each image may be represented as a collection of pixels that have values of red, blue, and green, each pixel may also be represented by the orthogonal dimensions of hue, saturation, and value (the last corresponding to the luminosity of the pixel). HSV histograms are often more subjectively informative than RGB histograms, as they are better at revealing effects such as tint, brightness, and color intensity Lee et al. (2005).

For this analysis, the histograms of hue, saturation, and value were obtained for each stimulus. Then, these stimuli were sorted the position of the peak in each distribution. This yields a 3D map of frequencies across stimuli, in which each row represents a single stimulus and each column represents a particular index in the histogram.

Figure 15 plots this multi-image histogram as a heat map for the hues of all photographic stimuli. In addition to the histograms for each individual image, Figure 15 also plots the marginal frequencies across all stimuli in each category. Here, we can see quite clearly that the different categories reliably have properties that can be used to distinguish one category from the next. Pictures of birds very frequently have green and cyan elements (because of leaves or sky), and flowers have a greater representation of yellow and purple. Photographs of people tend to be more reddish, while cats tend to be more orange.

Note that the apparently "blank bands" visible in these heat maps are black-and-white images. Since a black-and-white image cannot reasonable be described as having a particular hue, the frequency distribution for those images were uniform.

Figure 16 plots the histogram for saturation of photographic stimuli, and here, too, patterns of differ visibly. Photographs of flowers are typically highly saturated, while photos of cats and birds tend to have low saturation. However, an examination of the distributions of individual stimuli suggest that there is an overall level of heterogeneity in most cases, as evidenced by the lack of consistent vertical bands in the heat maps.

Figure 17 plots the histogram for value (i.e. brightness) of photographic stimuli, showing clear differences. Flowers and people tend to be spread across the range, while birds and cats tend to cluster toward the center. As in the case of saturation, there is a great deal of variation across stimuli, such that these would not be strongly selective signals.
Figure 13. Kernel density estimates of pixel entropy in the four photographic categories. The median image is indicated by the solid blue line, while the first and third quartiles are indicated by the blue dashed lines.
Figure 14. Kernel density estimates of pixel entropy in the four photographic categories. The median image is indicated by the solid blue line, while the first and third quartiles are indicated by the blue dashed lines.
Figure 15. Hue histograms for each photographic stimulus category, sorted by peak frequency. The histogram at the bottom depicts the marginal frequency across all images.
Figure 16. Saturation histograms for each photographic stimulus category, sorted by peak frequency. The histogram at the bottom depicts the marginal frequency across all images.
Figure 17. Value histograms for each photographic stimulus category, sorted by peak frequency. The histogram at the bottom depicts the marginal frequency across all images.
Figure 18. Hue histograms for each painterly stimulus category, sorted by peak frequency. The histogram at the bottom depicts the marginal frequency across all images.
**Figure 19.** Saturation histograms for each painterly stimulus category, sorted by peak frequency. The histogram at the bottom depicts the marginal frequency across all images.
Figure 20. Value histograms for each painterly stimulus category, sorted by peak frequency. The histogram at the bottom depicts the marginal frequency across all images.
Figure 18 plots the histogram for hue of the painting stimuli, and a much greater degree of uniformity is observed here than in the photographic stimuli. In all cases, painters favored colors in the orange-yellow and cyan ranges. This similarity across painters, combined with the heterogeneity of the images (in which many had no blue to them at all) ensures that hue could not be used as a reliable cue for differentiating between painters.

Figure 19 plots the histogram for saturation of the painting stimuli. Here, heterogeneity dominated, with all painters having images that were spread across the full range of saturations. While there were some differences (Dalí tended to be the most likely to have highly saturated colors, for example), the spread across the range prevented saturation from being a reliable cue.

Figure 20 plots the histogram for value of the painting stimuli. As in the case of saturation, the painters were highly heterogeneous, tending to favor intermediary values. This is unsurprising, as artists routinely avoid using pure white and pure black, instead favoring intermediate values that give an impression of contrast Escher (1989).

APPENDIX B: RHESUS MACAQUE TRAINING

In order to learn the stimulus categories, subjects were trained using a variant of the Concept Chain task. Rather than vary every image during every trial, only a subset of images initially varied. Trials otherwise resembled SimChain: Subjects were rewarded if they touched each image in the correct order. For example, during the first stage of training, the first three stimuli in the chain remained fixed for the duration of the session, but the fourth stimulus varied. For brevity, we will denote each stage of training with a code that indicates which stimuli remained fixed during a session and which varied. “F” denotes a fixed stimulus, and “V” denotes a varying stimulus. Thus, a four-item list in which the first three images are fixed but the fourth varies would be denoted by 1F-2F-3F-4V. Training sessions always lasted 40 trials, and each subject always maintained the prescribed order of categories indicated in section 2.2 of the main manuscript.

Training consisted of 15 stages in total. The first four stages varied only one image category at a time (Stage 1: 1F-2F-3F-4V; Stage 2: 1F-2F-3V-4F; Stage 3: 1F-2V-3F-4F; Stage 4: 1V-2F-3F-4F). The next six stages varied two categories at a time (Stage 5: 1F-2F-3V-4V; Stage 6: 1F-2V-3F-4V; Stage 7: 1V-2F-3F-4V; Stage 8: 1F-2V-3V-4F; Stage 9: 1V-2F-3V-4F; Stage 10: 1V-2V-3F-4F). The next four stages varied three categories at a time (Stage 11: 1F-2V-3V-4V; Stage 12: 1V-2F-3V-4V; Stage 13: 1V-2V-3F-4V; Stage 14: 1V-2V-3V-4F). Finally, during the last stage of training, all categories varied, making it a full-blown Category Chain (Stage 15: 1V-2V-3V-4V). Subjects advanced from one stage to the next according to a performance criterion (described below). Once the criterion was met for the final stage, an additional 25 sessions of data were collected. The results reported in section 4 are based on those 25 sessions.

Subjects learning the photographic categories advanced to the next stage of training when an 80% criterion was met. Subjects learning the painting categories displayed considerably more difficulty with the initial training, which necessitated two changes. First, criterions for advancing to the next stage was lowered to 70% throughout the training. Secondly, the first four stages of training (when only one category varied) were further subdivided into substages that had differing degrees of variation. Rather than varying across the entire image bank, the varying category initially only varied among two items. Once the criterion was met, the number of stimuli used in the varying category was increased to five. This proceeded over successive stages (5, then 10, then 25, then 50, then 100). Once subjects reached the fifth stage of training, these substages were no longer employed.
Figure 21. Estimated population distributions for human performance on the Category Chain. Gray violin plots represent the posterior population distribution for each parameter, whereas white violin plots represent the credible interval for the mean of those distributions. Top Row. Learning rate parameters during phase 1, using photographic stimuli. Bottom Row. Learning rate parameters during phase 2, using painting stimuli.

APPENDIX C: HUMAN LEARNING MODEL PARAMETERS

Parameter estimation for SimChain learning is complicated by two features of the model. The first (a problem in all non-linear regression regimes) is that the four learning parameters covary with one another. For example, if the twist parameter $v$ is doubled, this has the effect of shifting the entire distribution to the left, forcing the peak parameter $m$ to reduce its value accordingly.

The second problem is that when performance is close to either the floor or the ceiling, it becomes impossible to differentiate between values for the other parameters. For example, if the learning rate is close to zero, then the twist and peak parameters could have an enormous range of values without substantively impacting the shape of the learning curve.

Although Stan (Carpenter et al., 2015) is capable of robust inference regarding multi-level non-linear models, it cannot provide precise information about undecidable parameter values. This can be seen in Figure 21, which displays the learning parameters using photographic categories (top row) and painting categories (bottom row). In both cases, the peak parameter $m$ and the twist parameter $v$ showed poor convergence, as a consequence of both parameter covariance and parameter undecidability when learning rates were low.

A clearer picture emerged in the case of a model of the learning rates, depicted in Figure 22, which showed much more stable convergence. In general, reaction times were initially slower for early list items than for late list items, as shown by the downward trend in the intercept parameters $k$. Participants also
Figure 22. Estimated population distributions for linear models of reaction times on the Category Chain. Gray violin plots represent the posterior population distribution for each parameter, whereas white violin plots represent the credible interval for the mean of those distributions. **Top Row.** Reaction time model for phase 1, using photographic stimuli. **Bottom Row.** Reaction time model for phase 2, using painting stimuli.

tended to respond more rapidly as time passed, as evidenced by distributions of slope parameters $b$ tending to be less than zero.