Behavioral/Cognitive

Inferential Learning of Serial Order of Perceptual Categories by Rhesus Monkeys (*Macaca mulatta*)

Natalie Tanner,[1] Greg Jensen,[2,3] Vincent P. Ferrera,[2,4] and Herbert S. Terrace[3,4]

[1]Columbia College, [2]Department of Neuroscience, [3]Department of Psychology, and [4]Department of Psychiatry, Columbia University, New York, New York 10027

Category learning in animals is typically trained explicitly, in most instances by varying the exemplars of a single category in a matching-to-sample task. Here, we show that male rhesus macaques can learn categories by a transitive inference paradigm in which novel exemplars of five categories were presented throughout training. Instead of requiring decisions about a constant set of repetitively presented stimuli, we studied the macaque's ability to determine the relative order of multiple exemplars of particular stimuli that were rarely repeated. Ordinal decisions generalized both to novel stimuli and, as a consequence, to novel pairings. Thus, we showed that rhesus monkeys could learn to categorize on the basis of implied ordinal position, without prior matching-to-sample training, and that they could then make inferences about category order. Our results challenge the plausibility of association models of category learning and broaden the scope of the transitive inference paradigm.

Key words: categorization; cognition; serial learning; transitive inference

Significance Statement

The cognitive abilities of nonhuman animals are of enduring interest to scientists and the general public because they blur the dividing line between human and nonhuman intelligence. Categorization and sequence learning are highly abstract cognitive abilities each in their own right. This study is the first to provide evidence that visual categories can be ordered serially by macaque monkeys using a behavioral paradigm that provides no explicit feedback about category or serial order. These results strongly challenge accounts of learning based on stimulus–response associations.

Introduction

Since the discovery that pigeons could be trained to peck at photographs that only contain people (Herrnstein and Loveland, 1964), an extensive literature has demonstrated an animals' ability to categorize a wide variety of stimuli: faces (Marsh and MacDonald, 2008), plants, animals (Roberts, 1996), man-made objects (Bhatt et al., 1988), and even paintings (Watanabe, 2013). Freedman and colleagues (2001) found that primates could categorize computer-generated, systematically morphed images of cats and dogs. Activation in the lateral prefrontal cortex correlated with stimulus category even when subjects were required to sort morphed stimuli into new categories (Freedman et al., 2001).

Although many studies have demonstrated that animals can categorize stimuli, relatively little work has been done showing how categories are used in other cognitive tasks. Can animals, for example, treat categories as though they were informative stimuli, signaling appropriate behavior in a cognitive task?

Categorical serial learning

Altschul et al. (2016) demonstrated that rhesus macaques cannot only distinguish between four simultaneously presented categories of stimuli, but that they can also learn their serial order using a variant of the simultaneous chain task (Terrace, 1984, 2005). This suggests that animals cannot only learn to identify categories but they can also process categories in the same way they can process single stimuli. That is, they applied judgments of list position to entire classes of stimuli.

The transitive inference (TI) paradigm provides another method for studying serial learning: the ability to learn the relative order of a set of items. TI has been demonstrated in many species, including monkeys (McGonigle and Chalmers, 1992), mice (Van der Jeugd et al., 2009; Silverman et al., 2015), rats (Davis, 1992), pigeons (Lazareva and Wasserman, 2006), crows (Lazareva et al., 2004), and even fish (Grosenick et al., 2007; for

Received Jan. 27, 2017; revised May 10, 2017; accepted May 16, 2017.

Author contributions: G.J., V.P.F., and H.S.T. designed research; N.T. and G.J. performed research; N.T. and G.J. analyzed data; N.T., G.J., V.P.F., and H.S.T. wrote the paper.

This work was supported by National Institutes of Health Grant MH081153 to V.P.F. and H.S.T. We thank Grant Spencer, Orly Morgan, David Freshwater, and Carolina Montes for assistance with data collection.

The authors declare no competing financial interests.

Correspondence should be addressed to either Dr. Greg Jensen or Dr. Herbert S. Terrace, Columbia University, 1190 Amsterdam Avenue, MC 5501, 406 Schermerhorn Hall, New York, NY 10027. E-mail: greg.guichard.jensen@gmail.com or terrace@columbia.edu.

DOI:10.1523/JNEUROSCI.0263-17.2017

review, see Vasconcelos, 2008; Jensen, 2017). The TI paradigm requires subjects to maintain a representation of the relative order of list items. Following training using only adjacent pairs, above-chance ordering of nonadjacent pairs demonstrates that subjects were capable of TI (McGonigle and Chalmers, 1977; Jensen et al., 2013). While an animal's ability to learn a TI task and to categorize are well established, their ability to do both simultaneously has yet to be shown. Here, we assess the ability of rhesus macaques to learn category membership of stimuli that change on every trial, even as they learn the order of those categories and make TIs about them.

Our Category TI task follows the format of a traditional TI procedure of training using adjacent pairs and testing using all pairs but does so using stimuli that change after every trial. Subjects used trial and error to learn the category order of stimuli belonging to five categories: birds, cats, flowers, people, and hoofstock. Each trial begins with the presentation of two randomly selected pictures, drawn from a pool of 1000 images for each of the five image categories. Because the images included a range of related species photographed under varying conditions, subjects had to rely on category membership rather than their memory of specific stimuli.

Subjects learned all of the categories while learning the list order. They had no prior exposure to categorization tasks or to any of the stimuli belonging to those categories. After subjects were tested for TI with one stimulus order, the same categories were trained again in a different order. Subjects had to learn to sort the five categories into four different orderings during the course of the experiment. Given the size of the stimulus sets and the lack of prior training, a demonstration of TI under these conditions would show that perceptual categories can be learned and represented in the same flexible fashion as the constant stimuli that are normally used to train TI. It would also show that subjects can learn to categorize images without first being trained regarding category membership (e.g., using match-to-sample).

Materials and Methods

Subjects. Subjects were two adult male rhesus macaques (*Macaca mulatta*), Subjects N and O. Subject N was 8 years old and had minimal experience with the TI procedures. Subject O was 22 years old and had extensive experience performing TI tasks that may have facilitated his learning of the Categorical TI task. Neither subject had any experience categorizing pictorial stimuli. Subjects' first exposure to these five categories began at the start of this experiment.

The subjects were individually housed in a colony room containing approximately two-dozen macaques and performed the experimental tasks in their home cage. The subjects were trained 5 d a week, one session each day. To increase their motivation to perform the task for fluid reward, the monkeys were put on fluid restriction (300 ml of water per day) 2 d before the first day of testing. Depending on task performance, subjects could earn up to 500 ml a day performing the task, with 200–300 ml being typical. Most days, subjects earned their entire fluid ration performing the task. This was supplemented as needed after the experimental session ended to meet the minimum requirement. Each monkey received a set amount of biscuits before testing. Fruit was distributed following testing.

The study was performed in accordance with the guidelines provided by the *Guide for the care and use of laboratory animals* of the National Institute of Health. This work, performed at the Nonhuman Primate Facility of the New York State Psychiatric Institute, was overseen by New York State Psychiatric Institute's Department of Comparative Medicine and was approved by the Institutional Animal Care and Use Committees at Columbia University and New York State Psychiatric Institute.

Apparatus. The apparatus used for this study was an in-cage testing device with a touch-screen tablet and a fluid delivery system comprising a 1 L calibrated reservoir and a solenoid valve. The solenoid valve was controlled by the tablet computer via an Arduino Nano interface. Each correct response delivered 0.25 ml of water through a spigot below the touchscreen. The entire testing device fit snugly and securely into the doors of the monkey's home cages. The tablet had a 10.1 inch HD display, operated at 1266 × 768 resolution, and used capacitative multitouch inputs. All tasks were programmed in JavaScript and run in a Google Chrome browser window under a Windows 8.1 operating system.

All stimuli used in the experimental tasks were 250 × 250 pixel images presented randomly on the left- and right-hand sides of the tablet's screen. Between trials, a solid blue square was presented at the center of the screen. Touching it initiated a new trial. This focused the subject's attention and directed the subject's hand toward the center of the screen to reduce response bias.

Stimuli. Stimuli were selected from five categories: birds, cats (including both housecats and large predatory cats), flowers, people, and hoofstock (the last being a mix of sheep, cows, horses, and goats). Other than people, each category comprised a variety of species photographed under varying conditions. For each category, subjects were exposed to 1000 different stimuli. It was therefore highly unlikely that subjects would see the same image more than once within an interval of several hundred trials. Stimuli from the first four categories were previously used by Altschul et al. (2016) with a different set of subjects.

The TI procedure. During training, subjects were provided with accurate but incomplete information about list order. It was, however, possible for them to infer the relative ordinal position of each item. Consider, for example, a list of arbitrarily selected stimuli (A–D, E) in which the order was determined by the experimenter and unknown to subjects. On each trial, subjects were presented with a pair of items. A response to the item from the earlier list position was always rewarded. If, for instance, the order was ABCDE and the pair BC was presented, a response to B was rewarded because it came first. If, however, the pair AB was presented, the subject had to choose A to receive a reward. Following training on adjacent items (AB, BC, etc.), the critical question is whether subjects were able to infer the correct choice when presented with nonadjacent items that they had never seen previously (e.g., AC).

Each session, subjects completed up to 1000 trials of TI training (Fig. 1) by touching stimuli on the tablet to earn water rewards. Each of two images presented during a trial had an associated "list rank" that was not explicitly communicated to the subjects. The image with the lower rank (i.e., earlier in the list) was always correct. Selecting the correct item yielded a reward of 0.25 ml of water. Image ranks ranged from 1 to 5. Thus, subjects were effectively asked to discover the order of a five-item list (denoted as ABCDE) by pairwise trial and error (Jensen et al., 2013, Jensen and Altschul, 2015) in a procedure in which the exemplars of each category were selected at random and seldom repeated.

Unlike traditional TI tasks, a particular rank was not associated with a single static image. Instead, as described earlier, rank was associated with a stimulus category. Every time a subject saw the pair AB, it consisted of a different random pair of images from categories A and B than those shown in the previous pairing of A and B. This meant that subjects could not solve the task by learning the order of specific stimuli. Because the images included a range of related species photographed under varying conditions, subjects had to generalize their understanding of one image of a bird and one image of a cat and understand, for example, that all birds come before all cats. Because subjects had no experience with these categories, they had to learn them at the same time they were learning list order. In this respect, our procedure deviated from the typical matching-to-sample or match-to-category procedures used to study concept-formation (Freedman et al., 2001; Bodily et al., 2008).

To test subjects' knowledge of TI, their initial training was limited to adjacent pairs (AB, BC, CD, DE). During such training, A was always rewarded, E was never rewarded, and all other stimuli were rewarded half the time. B, for example, was correct when paired with C, but incorrect when paired with A. Its asymptotic expected value was therefore 0.5. Once subjects performed above chance on such pairings, they were tested on the "critical test pair," BD. Because B and D each has an expected value of 0.5, associative models predict performance no better than chance.

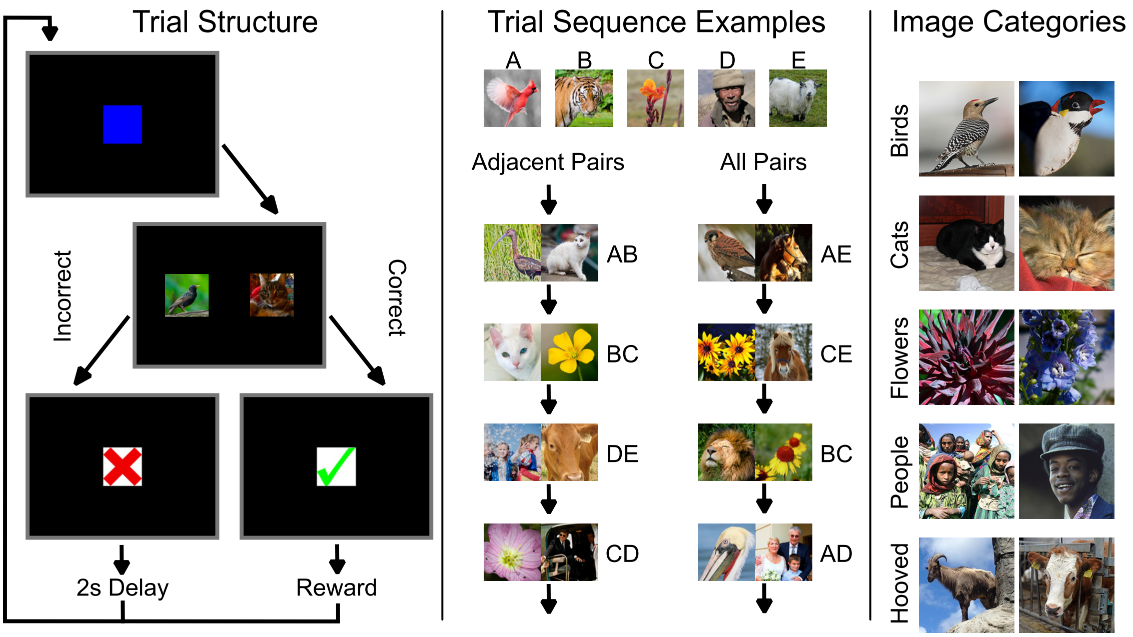

Figure 1. Procedure for the categorical TI task. Left, Structure for any single trial of the task. Subjects must touch a blue square to begin the trial, which is immediately replaced by two images. If a correct response is made, subjects see a green checkmark and are immediately given a fluid reward. If an incorrect response is made, subjects see a red X, followed by a black screen for 2 s. Following feedback, the next trial begins with the start stimulus. Middle, Each phase of the experiment made use of a consistent category sequence (in this case, birds-cats-flowers-people-hooved). The stimuli themselves, however, were drawn at random from the image bank during every trial. During adjacent-pair trial (using only AB, BC, CD, and DE), the identity of the stimulus changed for every trial, even when the same category appeared in two consecutive trials. The left-right position of stimuli was also counterbalanced. This was also the case during all-pairs sessions, which intermixed all possible stimulus pairings. Right, Two exemplars each from the five stimulus categories used in the experiment. In all categories, an effort was made to include category members from multiple distances and angles, with a mixture of both solitary and group photos, as well as both color and black-and-white. This stimulus diversity was intended to reduce subjects' reliance on specific discrete features as category cues. The individual stimulus images are reproduced under Create Commons licenses.

Contrary to this prediction, subjects across many species routinely favor B, thereby displaying TI, despite B and D having similar reward values. Actual reward history may differ from this asymptotic value, contingent on subject performance.

After at least six sessions of adjacent pair training, subjects were exposed to all 10 possible stimulus pairings. Knowledge of TI would be demonstrated if subjects performed above chance on the critical pair BD.

The symbolic distance effect is a robust feature of TI performance. Stimulus pairs that are more widely separated in the list show higher levels of accuracy than those that are closer together (D'Amato and Colombo, 1990; Treichler et al., 2007). Given our "train-adjacent-test-nonadjacent" task design, a symbolic distance effect observed at the start of each all-pairs block of sessions would be difficult to explain using associative models, as would above-chance performance on critical test pairs. Such effects are instead consistent with a strategy that relies on the comparison of relative ordinal or spatial position, as it should be easier to discriminate between widely-spaced items than closely-spaced items.

After an initial transfer from adjacent pairs to all pairs with respect to a particular list order, subjects were again presented with adjacent pairs, this time using a different ordering of categories. They repeated the adjacent-pair-training, all-pair-testing design for three more phases, yielding a total of four different category orders. The order of the categories used for Phase 1 was ABCDE, with BD being the key pair for evidence of TI. The order for Phase 2 was DBCEA, with BE being the key pair. The order for Phase 3 was AECBD, with EB being the key pair. The order for Phase 4 was EDCBA, with DB being the key pair. Because of scheduling conflicts and technical difficulties, subjects were not run for the same number of sessions: Subject N completed 80 sessions, whereas Subject O completed 60 sessions. Both subjects consistently completed three sessions before and after each transition.

Statistical analysis. Behavior was modeled using logistic regression, building on the method described Jensen et al. (2013). The probability of selecting the correct stimulus on trial during a particular session is given by $p(t)$, which was fit according to the following function:

$$p(t) = (1 + \exp(-(\beta_\phi + \beta_t t + \beta_D(D - 2.5) + \beta_{Dt}(D - 2.5)t)))^{-1} \quad (1)$$

Here, t refers to the trial number, beginning with zero; consequently, β_ϕ is the intercept term, and β_t is the slope as a function of time. D refers to the symbolic distance between the list positions of the stimuli (e.g., for an adjacent pair). Because the maximum value of D is 4 (in the case of pair AE), subtracting 2.5 from D in the analysis centers the results with respect to distance. As a result, β_D provides an estimate of improvement in performance overall and represents the differential performance that results from the symbolic distance effect. β_{Dt} represents the interaction between overall learning and the symbolic distance effect. The logistic function provides a more compressed version of Equation 1 as follows:

$$p(t) = \text{logistic}\ (\beta_\phi + \beta_t t + \beta_D(D - 2.5) + \beta_{Dt}(D - 2.5)t) \quad (2)$$

A different logistic regression was performed for each subject during each session because subjects were presented with the same stimuli over multiple consecutive sessions. This allows us to distinguish between behavior during learning and behavior when performance reached ceiling (which, in macaques, is consistently below perfect accuracy).

Reaction time was also evaluated on a per-session basis, given a log-linear model as follows:

$$\log\ (\text{reaction time}) = \gamma_\phi + \gamma_D(D - 2.5) \quad (3)$$

Because D was centered with respect to symbolic distance, the intercept γ_ϕ can be interpreted as the mean of the log reaction times, whereas γ_D is responsible for the deviation as a function of distance. This model was fit for each subject during each session.

Models were fit using the Stan language (Carpenter et al., 2017). To facilitate continuity from one session to the next, model estimates for a subject's performance at the end of each session acted as a regularizing prior on that subject's performance at the beginning of the following

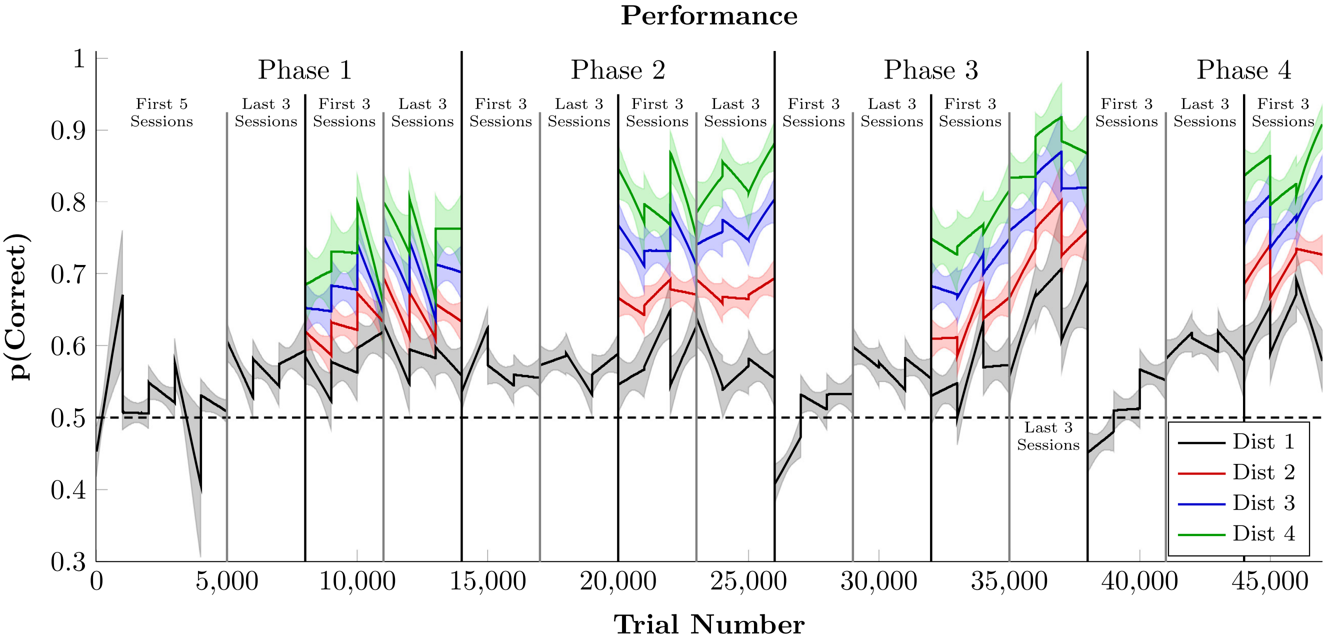

Figure 2. Time series analysis of task performance, divided by symbolic distance, averaged across subjects. All sessions presented adjacent pairs (black), but only all-pairs sessions included symbolic distance of 2 (red), 3 (blue), and 4 (green). Discontinuities correspond to gaps between sessions. Shaded regions represent the 99% credible interval of the estimate.

session. In transitions between phases, earlier performance was not used as a prior.

In keeping with the spirit of the Stan language, we did not perform null-hypothesis significance tests. As articulated by Gelman (2013), we do not feel that *p* values reported with a "discovery" mindset are appropriate to the data at hand. We are instead interested in reporting the relevant uncertainty about the parameters, in keeping with a "measurement" mindset. Throughout, we report 80% and 99% credible intervals for estimates of parameter values and task performance. When a so-called "null result" (e.g., a parameter of zero) is omitted from these credible intervals, the Bayesian interpretation of the interval is that, given the model, the data, and our prior assumptions, our conclusion is that it is a very unlikely value for the parameter to take. In this respect, we are more interested in measuring the size of the symbolic distance effect, and less interested in discovering whether it differs from the null.

Results

We achieved both of our goals by showing that rhesus macaques could, during TI training, learn (1) to simultaneously categorize photographs from five categories without prior matching-to-sample training and (2) the ordinal position of those categories in an implicitly defined list.

Figure 2 presents the estimated mean probability of a correct response (combining the uncertainty from both subjects) during the first and last three sessions of each phase. Adjacent pairs are plotted in black, while the 80% credible interval for those estimates is shown by the shaded regions. Consistent with the past literature (Terrace et al., 2003; Lazareva and Wasserman, 2006; D'Amato and Colombo, 1990), performance on adjacent pairs was above chance but comparatively low.

After at least 6 sessions of adjacent-only training, monkeys were tested with all category pairings. Nonadjacent pairs yield higher accuracy, even in the earliest trials of each all-pairs phase. A symbolic distance effect is clearly visible. Response accuracy was highest for the largest symbolic distance of 4 (depicted in green), whereas distance 3 (in blue) and 2 (in red) yielded intermediate response accuracies. Although exemplars changed on every trial, a distance effect appeared immediately after the transition from an adjacent-pair to an all-pair design. This suggests that subjects immediately made TIs at the category level. Figure 2 incorporates data from 47 sessions for each subject.

Figure 3 (left) presents a more direct depiction of the distance effect from the logistic regressions using the mean estimated parameter, measured in log units. This estimates the differential impact of symbolic distance, independent of overall performance. A positive value for the parameter indicates the traditional distance effect, with larger values corresponding to more dramatic effects. For example, if (hypothetically) accuracy on adjacent pairs was at chance (i.e., $\frac{1}{1 + \exp(0.0)} = 0.5$ accuracy) and if $\beta_D = 0.3$, then accuracy on a pair with a distance of two would be $\frac{1}{1 + \exp(-0.3)} = 0.574$, and distance 3 would be $\frac{1}{1 + \exp(-0.6)} = 0.646$, and so forth. In all but three of the sessions, the 99% credible interval of the mean (depicted by the whiskers) excludes zero. The 80% credible interval (depicted by the boxes) exclude zero for all sessions.

In the evaluation of TI, the test pair BD is particularly important because it is the only nonadjacent pair for which (after adjacent-pair training) both stimuli have an expected value of 0.5. Performance above chance on this pair is generally taken to be evidence that an inference has taken place. Figure 3 (right) provides three estimates for each subject of the probability of a correct response on the pair BD at transfer for each subject. The first estimate (labeled "trials only") is based only on the first two presentations of BD in each phase. Both subjects chose B more often than they did D (6 of 8 times for Subject N and 5 of 8 times for Subject O). Because there were so few transfers, however, the uncertainty associated with these estimates is large. The second estimate (labeled "BD model fit") estimates the probability of a correct response based on a logistic regression using only BD trials across all phases. Finally, the third estimate (labeled "full model fit") uses Equation 1 to make an estimate using all trials to infer an estimate for BD. Because the second and third estimates incorporate other data as a time series, they can estimate mean

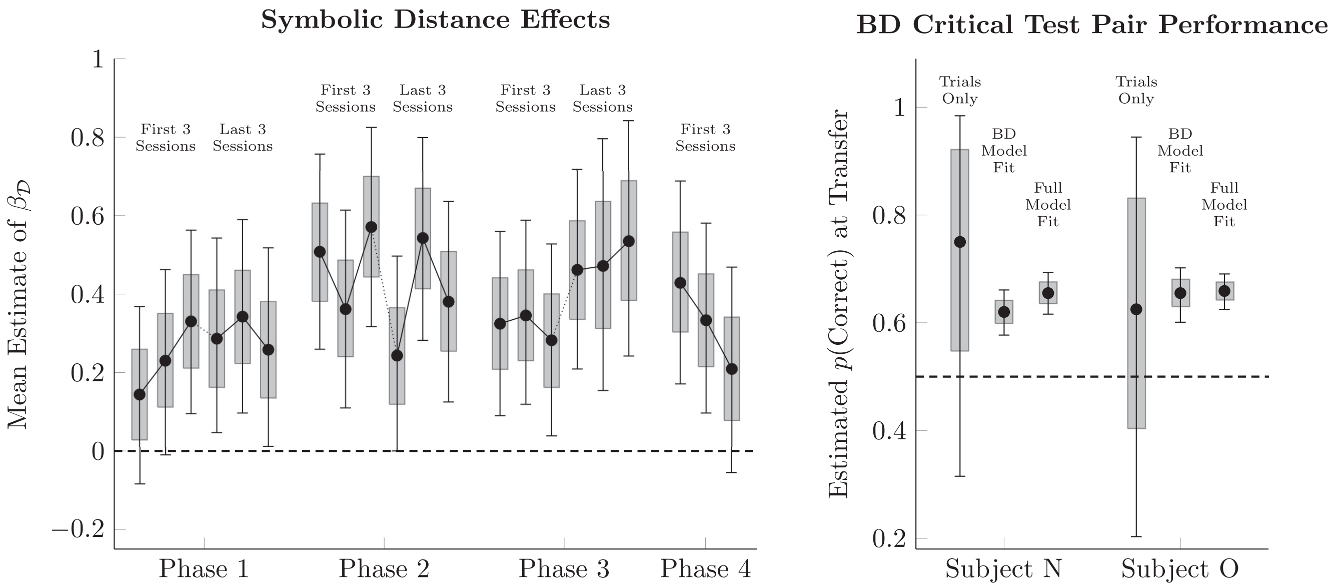

Figure 3. Evidence of distance effects at transfer. Whiskers represent 99% credible intervals for the estimates. Shaded intervals represent 80% credible intervals. Left, Session-by-session of the "distance effect on trial zero" parameter in the logistic regression analysis of performance (β_D in Eqs. 1, 2) during all-pairs sessions, averaged across subjects. Because parameters are measured in log-odds units, no distance effect at transfer would correspond to a parameter value of 0.0. Right, Proportion of correct responses for the critical test pair BD on its first presentation. "Trials only" estimates are based only on the first two BD presentations in each phase. "BD model fit" estimates are based on the intercept of a logistic regression of response accuracy that uses only BD trials. "Full model fit" estimates use Equation 1 to predict BD accuracy using all trials and their symbolic distances. Performance above chance indicates that TI has occurred.

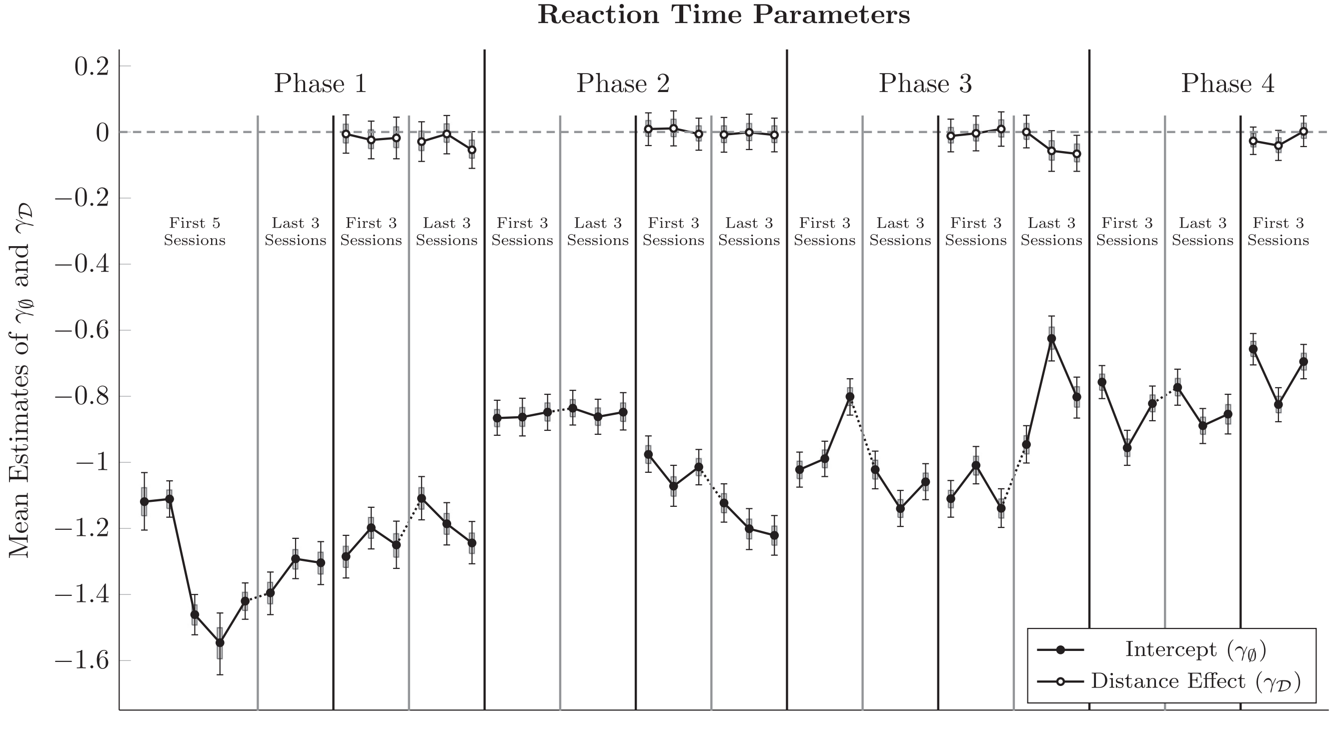

Figure 4. Session-by-session of the intercept parameter (γ_ϕ in Eq. 3, in black) and "distance effect on trial zero" parameter (γ_D in Eq. 3, in white) in the regression analysis of log reaction time, averaged across subjects. Values of near zero indicate no differential effect on reaction time as a function of symbolic distance. Whiskers represent 99% credible intervals for the estimates. Shaded intervals represent 80% credible intervals.

performance at transfer with greater confidence, leading to credible intervals that reliably exclude chance performance.

Figure 4 plots mean parameter values across subjects for both the intercept and the distance effect. Overall, reaction time increased during successive phases of the experiment, from a minimum group mean of −1.55 log seconds (0.21 s) to subsequent values reliably exceeding −0.8 log seconds (0.45 s). This suggests that subjects became less impulsive with training. Thus, despite a reliable distance effect for response accuracy, none was obtained for reaction time.

It is noteworthy that subjects continued to make errors at a high rate after extensive training on all pairs, even on "easy" pairs like AE. Monkeys given extensive training on 5 item lists are ordinarily able to respond correctly to the pair AE with close to no

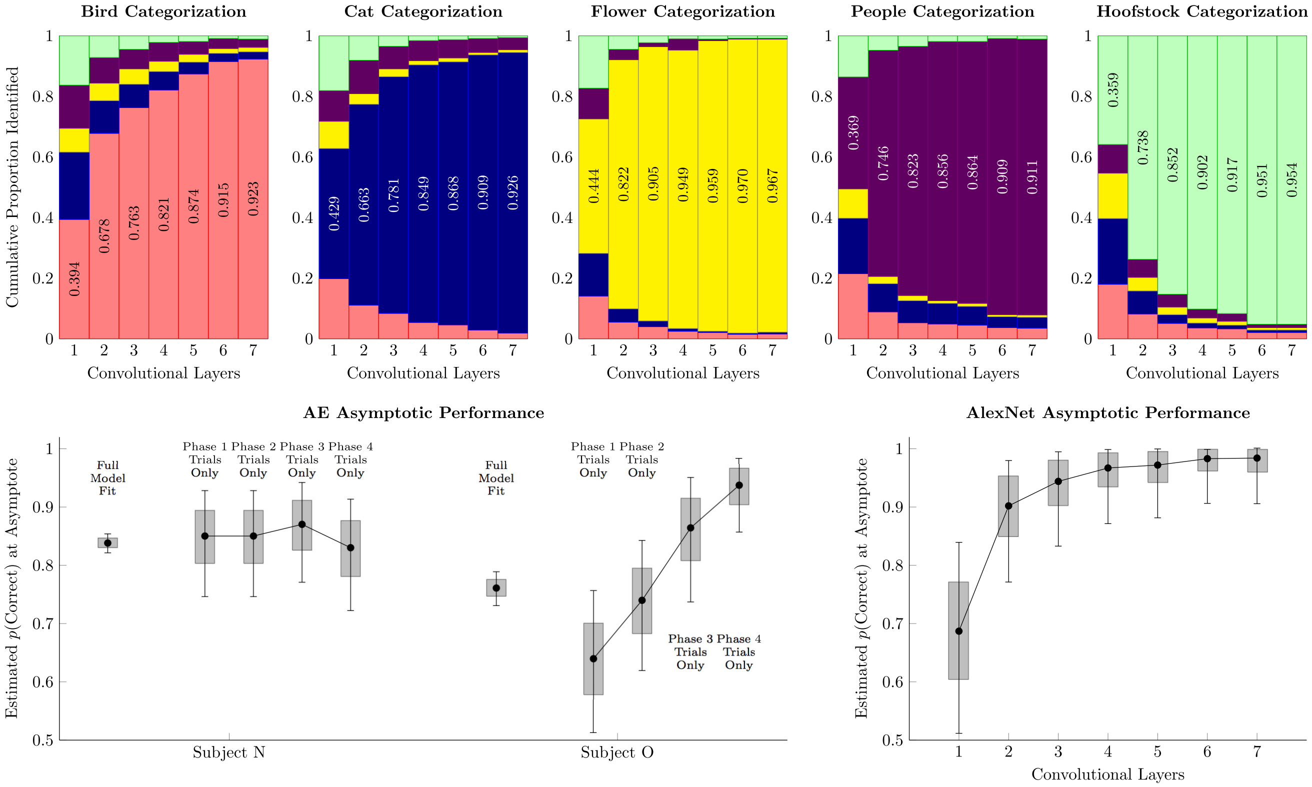

Figure 5. Comparison of monkey performance and AlexNet, a deep learning network trained for image classification. Whiskers represent 99% credible intervals for the estimates. Shaded intervals represent 80% credible intervals. Top, Image classifications for each of the five categories by AlexNet, given the outputs at each of its seven convolutional layers. Each block assumes that a stimulus of the type given by the title has been presented, and cumulative proportions of the five classified categories are (plotted from bottom to top) birds (light red), cats (dark blue), flowers (yellow), people (dark purple), and hoofstock (light green). The proportion of correct classifications is also inscribed in the relevant block. Bottom left, Subjects' response accuracy for the pair AE at the end of each phase. "Full model fit" estimates use Equation 1 to estimate performance based on all trials across all phases. "Trial only" estimates are based on the proportion of correct responses to AE during the last session of each phase. Bottom right, AlexNet classification accuracy in forced two-item classification as a function of the number of convolutional layers used. The uncertainty in the estimates arises across multiple simulations using different training and validation sets of stimuli.

errors (D'Amato and Colombo, 1988; Treichler and Van Tilburg, 1996). The failure to do so after thousands of trials of training likely reflects the greater difficulty introduced by stimulus categorization. Characterizing how difficult the task is, however, requires an objective benchmark against which to compare performance. To assess this, we used AlexNet, a deep learning network (DLN) trained for image classification (Krizhevsky et al., 2012). Like other DLN classifiers, AlexNet relies on a convolutional hierarchy, wherein a single node at one layer received weighted inputs from large numbers of nodes at the preceding layer. These weights were previously trained to act as feature detectors, such that, although the inputs are sensitive only to brightness and color, the first convolutional layer can be sensitive to contours and simple shapes. With each additional convolutional layer, more and more sophisticated features can be built up from those at lower layers. For example, in a DLN optimized to detect faces, nodes in the second convolutional layer can readily pick out patterns that resemble eyes and mouths, whereas nodes in the third convolutional layer can pick out patterns that resemble complete faces (Lee et al., 2009). Of the 23 layers in AlexNet, 7 are convolutional and have previously been optimized to classify images into 1 of 1000 different categories. Thanks to this feedforward approach, DLN models, such as AlexNet, bypass the debate over what constitutes a "discrete feature" by defining features in terms of statistical regularities within each category.

By performing image classification using only the first *x* convolutional layers of the network, we can effectively handicap performance by allowing only a particular level of feature-like regularity to be discovered. We can then compare this range of response accuracies to that of the monkeys. Because AlexNet was not originally trained on our stimuli, we retrained it using random 300-image subsets as training sets and the remaining 700 images as validation sets. This gave a range of performance accuracies, which can be interpreted as revealing how higher and higher levels of feature abstraction can be inferred by the network and used to improved performance.

Figure 5 (top) shows the mean probabilities of how AlexNet is likely to classify a stimulus given some number of convolutional layers. For example, when presented with a picture of a bird, AlexNet's first convolutional layer only identified it as a bird 39.4% of the time (instead confusing it for a cat 22.2% of the time). However, as convolutional layers were added, image classification steadily rose, reaching 92.3% when all seven layers were included. Using AlexNet as a benchmark, photographs of flowers were generally easiest to distinguish from the other image categories (96.7% accuracy using the full network), whereas photographs of people were the most difficult to classify (91.1% accuracy using the full network).

Figure 5 (bottom right) shows the probability that AlexNet could correctly identify each stimulus in a forced choice between pairs. These results take into account the possibility that the classifier would have to guess. If, for example, only the first layer of the network was presented with a bird and a horse, and it knew that "bird" was the correct answer, it would choose the bird

68.7% of the time on average because a single-layer network has some difficulty telling birds from horses. However, the full seven layers would choose the bird 98.4% of the time on average.

We can compare these benchmarks to the performance of the subjects on pair AE, shown in Figure 5 (bottom left). These represent performance at the end of each phase, averaged over the final session of all-pairs TI. The "full model fit" shows estimated response accuracy for AE using Equation 1 across all phases. The "trials only" fits for each phase are mean accuracy based only on the last session of each phase. Subject N had slightly higher asymptotic accuracy on AE trials than Subject O overall. However, Subject O showed steady improvement from one phase to the next. Compared with AlexNet, Subject N reliably performed at a level comparable with the first two layers of AlexNet, whereas Subject O began at a level comparable with only a single layer, and ended at a level comparable with three or four layers.

Discussion

Unlike earlier studies of category formation, we showed that rhesus macaques could be trained by a TI paradigm to differentiate five perceptual categories (birds, cats, flowers, people, hoofstock) and to learn their ordinal positions on four different implicit lists. Remarkably, the vast majority of stimulus pairs were trial-unique.

Although the estimated distance effect was consistently positive (i.e., larger symbolic distances yielded higher performance), parameter estimates for the first two sessions of Phase 1 did not rule out a value of 0.0. During those sessions, subjects may still have been learning to categorize the exemplars or may have had a less robust understanding of the category ordering. However, in every subsequent transfer, subjects showed a clear distance effect. Together, our results show that monkeys could retain knowledge of five distinct perceptual categories, despite changes to the ordering of the categories. They could readily update the ordering of those categories and improve their performance with experience.

Past studies have shown that monkeys' performance improves as they accrue expertise over consecutive sessions learning serial tasks (Terrace et al., 2003). Response accuracy in later stages resembled TI in other studies (e.g., Jensen et al., 2015), suggesting that given sufficient expertise, subjects were able to manipulate categories as though each was a "stimulus." However, despite extensive training during each phase, asymptotic performance still included many errors. Compared with AlexNet, subjects performed in a fashion similar to a two- or three-layer DLN. This is consistent with other reports (e.g., Cadieu et al., 2014) that DLN image classification performance now exceeds that of nonhuman primates.

The analysis of reaction times yielded two surprising results. Extensive training increased reaction time and, unlike response accuracy, no reliable distance effect of reaction time emerged. These effects are likely due to the change of exemplars on every trial. Whereas a monkey that lacks category knowledge can respond rapidly by guessing, a monkey that seeks to classify an exemplar may need more time to identify it. The lack of a distance effect is consistent with previous work on categorical decision-making in which it has been found that animals have a fixed, short response time that does not vary with decision difficulty (Uchida et al., 2006).

Traditionally, studies of categorization in animals initially train category membership using the match-to-sample paradigm (Herrnstein and Perrett, 1985; Crouzet et al., 2012), a match-to-stimulus paradigm (Fabre-Thorpe et al., 1998; Basile and Hampton, 2013), or a match-to-category design (Freedman et al., 2001). In these paradigms, subjects evaluate stimuli one at a time, a process that is vulnerable to a "guessing" strategy (Jensen and Altschul, 2015). The categorical TI experiment is distinct from these procedures because it requires subjects to evaluate two categories at a time in 10 possible pairings. Subjects not only learned to discriminate the categories, but did so while simultaneously learning the ordinal positions of those categories.

Neural substrates of categorization and serial order

Studies in humans and nonhuman primates have identified several cortical regions with activity related to visual categorization and serial order in a variety of behavioral paradigms. In monkeys, activity that is informative about visual category boundaries has been reported in multiple regions of parietal and prefrontal cortex, including the lateral intraparietal area (Freedman and Assad, 2006), dorsolateral prefrontal cortex (Freedman et al., 2001), inferotemporal cortex (Freedman et al., 2003), and frontal eye field (Ferrera et al., 2009) (for review, see Freedman and Assad, 2016). Similar regions in prefrontal and inferotemporal cortex contain neurons that respond selectively when monkeys engage in tasks that require explicit or implicit ordering of visual images (Miyashita, 1988; Berdyyeva and Olson, 2010; Brunamonti et al., 2016).

Human studies using fMRI have likewise implicated prefrontal and parietal cortex in TI tasks, as well as the hippocampus (Heckers et al., 2004; Van Opstal et al., 2008; Goel et al., 2009; Zalesak and Heckers, 2009; Wendelken et al., 2010;Koscik et al., 2012). Activity related to visual categories has been identified in human inferotemporal cortex (Mur et al., 2012). Together, the human and monkey studies point to a network of prefrontal, parietal, and inferotemporal cortical regions involved in both categorization and serial learning, raising the possibility that these capabilities might be colocalized. So far, no studies have reported a neural representation for the serial order of visual categories.

Cognitive representation of serial categories

Proposals of how animals categorize stimuli can be grouped into two classes: associative learning and cognitive representation. Roberts (1996) and Lea and Ryan (1984) argued that animals' ability to categorize can be explained by their reinforcement history. Because category exemplars contain particular features, they can be paired with rewards. But this interpretation raises an obvious question: What are those features? Herrnstein and Perrett (1985) questioned that interpretation in an experiment in which photographic stimuli were randomly assigned to categories without regard to their content. Pigeons were nevertheless able to learn which images belonged to which category.

A more modern cognitive approach treats a perceptual category as a "conceptual representation" (Newen and Bartels, 2007). Under such a view, categorization arises from an animal's ability to embed stimuli into a representational hierarchy, such that stimuli can both be decomposed into features and also be grouped into categories. These groupings can be defined statistically, rather than by strict rules. For example, although "humans" usually have two eyes, an animal would recognize someone with only one eye as human if enough other features were consistent with that label. Thus, no single feature is necessary or sufficient to determine category membership. Instead, a hierarchical inference permits categorization. Such representations fall short of the abstraction of language but are more flexible than reward associations to features. Furthermore, "features" themselves can also be defined hierarchically, as is the case for DLNs, such as AlexNet.

This approach is already being used to better understand the sensory cortex (Yamins and DiCarlo, 2016), and so may therefore provide a framework for understanding categorization more broadly.

Before the 1970s, TI was thought to rely on logic, limiting it to humans old enough to possess both language and a capacity for concrete operations (Vasconcelos, 2008). However, Bryant and Trabasso (1971) demonstrated that 4-year-old children displayed TI before the manifestation of concrete operations, suggesting a more fundamental mechanism. Their trial-and-error method of training was translated to nonhuman animals by McGonigle and Chalmers (1977), who found evidence of TI in squirrel monkeys (*Saimiri sciureus*).

Although evidence for TI in animals is compelling, its underlying mechanism remains unclear. Some have argued that associative learning (often using some variant of the Rescorla–Wagner model) is the most parsimonious explanation for TI's ubiquity in animals (Vasconcelos, 2008). However, there are serious problems with this argument. According to associative models, the massed presentation of a single stimulus pair (e.g., DE) should bias responding toward the correct item in that pair, even in pairings where it is incorrect (e.g., BD). However, several species have demonstrated robust response accuracy despite these manipulations (Lazareva and Wasserman, 2012; Jensen et al., 2017).

Both the consistent manifestation of symbolic distance effects and TI's resistance to the effect of massed trials suggest that behavior is mediated by cognitive representations, updated based upon feedback. For example, Jensen et al. (2015) proposed a Bayesian model in which subjects estimate the spatial positions of stimuli along an abstract continuum and perform inferences by sampling from those uncertain estimates. Our symbolic distance effects and transfer effects for critical test pairs are consistent with the predictions of a Bayesian spatial model.

That said, it would be a mistake to make too broad a claim, based on our data, about categorization and TI. All of our stimuli were photographs. Despite a variety of angles, colors, and degrees of zoom, there inevitably were statistical regularities among images. Pictures of flowers never included eyes, so their absence could be used as a cue for that category. We do not rule out the possibility that subjects relied on a classifier that was tailor-made for the stimulus set, shaped by the task's feedback (Jensen and Altschul, 2015). However, this does not alter our conclusions regarding serial learning. A tailor-made classifier might perform more poorly on novel stimuli, but even then, the subjects would be performing TI at a level of abstraction above that of specific stimuli.

Another potential concern regarding photographic stimuli is that they may be "ecologically relevant," such that subjects might be biologically predisposed to categorize them correctly (New et al., 2007). A replication of our design using artificial stimuli (e.g., man-made stimuli) would be illuminating. However, we make no claims about how categorization is performed, or whether it is innate or acquired. Past studies of animal categorization suggest that animals exhibit serial learning with abstract artificial stimuli (Altschul et al., 2016), and that they can categorize visually degraded photographic stimuli (Basile and Hampton, 2013) and artificial stimuli (Matsukawa et al., 2004). Although our own use of photographic stimuli may introduce an ecological confound, an ample literature suggests that subjects should be able to learn to categorize and serially order stimuli beyond those that are "ecologically relevant."

Notes

Supplemental material for this article is available at https://github. com/vpflab/transitive-inference-experiment/tree/master/CategoricalTIStudy. Repository for data analysis script and raw data. This material has not been peer reviewed.

References

Altschul D, Jensen G, Terrace HS (2016) Perceptual category learning of photographic and painterly stimuli in rhesus macaques (*Macaca mulatta*) and humans. PeerJ PrePrints 4:e967v3.

Basile BM, Hampton RR (2013) Monkeys show recognition without priming in a classification task. Behav Process 93:50 – 61. CrossRef Medline

Berdyyeva TK, Olson CR (2010) Rank signals in four areas of macaque frontal cortex during selection of actions and objects in serial order. J Neurophysiol 104:141–159. CrossRef Medline

Bhatt RS, Wasserman EA, Reynolds WF Jr, Knauss KS (1988) Conceptual behavior in pigeons: categorization of both familiar and novel examples from four classes of natural and artificial stimuli. J Exp Psychol Anim Behav Process 14:219 –234. CrossRef

Bodily KD, Katz JS, Wright AA (2008) Matching-to-sample abstract-concept learning by pigeons. J Exp Psychol Anim Behav Process 34:178 – 184. CrossRef Medline

Brunamonti E, Mione V, Di Bello F, Pani P, Genovesio A, Ferraina S (2016) Neuronal modulation in the prefrontal cortex in a transitive inference task: evidence of neuronal correlates of mental schema management. J Neurosci 36:1223–1236. CrossRef Medline

Bryant PE, Trabasso T (1971) Transitive inference and memory in young children. Nature 232:456 – 458. CrossRef Medline

Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10:e1003963. CrossRef Medline

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A (2017) Stan: a probabilistic programming language. J Stat Softw 76:1–32.

Chen S, Swartz KB, Terrace HS (1997) Knowledge of the ordinal position of list items in rhesus monkeys. Psychol Sci 8:80 – 86. CrossRef

Crouzet SM, Joubert OR, Thorpe SJ, Fabre-Thorpe M (2012) Animal detection precedes access to scene category. PLoS One 7:e51471. CrossRef Medline

D'Amato MR, Colombo M (1988) Representation of serial order in monkeys (*Cebus apella*). J Exp Psychol Anim Behav Process 14:131–139. CrossRef Medline

D'Amato MR, Colombo M (1990) The symbolic distance effect in monkeys (*Cebus apella*). Anim Learn Behav 18:133–140. CrossRef

Davis H (1992) Transitive inference in rats (*Rattus norvegicus*). J Comp Psychol 106:342–349. CrossRef Medline

Fabre-Thorpe M, Richard G, Thorpe SJ (1998) Rapid categorization of natural images by rhesus monkeys. Neuroreport 9:302–308. Medline

Ferrera VP, Yanike M, Cassanello C (2009) Frontal eye field neurons signal changes in decision criteria. Nat Neurosci 12:1458 –1462. CrossRef Medline

Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. Nature 443:85– 88. CrossRef Medline

Freedman DJ, Assad JA (2016) Neuronal mechanisms of visual categorization: an abstract view on decision making. Annu Rev Neurosci 39:129 – 147. CrossRef Medline

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312–316. CrossRef Medline

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J Neurosci 23:5235–5246. Medline

Gelman A (2013) Interrogating *p*-values. J Math Psychol 57:188 –189. CrossRef

Goel V, Stollstorff M, Nakic M, Knutson K, Grafman J (2009) A role for right ventrolateral prefrontal cortex in reasoning about indeterminate relations. Neuropsychologia 47:2790 –2797. CrossRef Medline

Grosenick L, Clement TS, Fernald RD (2007) Fish can infer social rank by observation alone. Nature 445:429 – 432. CrossRef

Heckers S, Zalesak M, Weiss AP, Ditman T, Titone D (2004) Hippocampal activation during transitive inference in humans. Hippocampus 14:153–162. CrossRef Medline

Herrnstein RJ, Loveland DH (1964) Complex visual concept in the pigeon. Science 146:549–551. CrossRef Medline

Herrnstein RJ, Perrett DI (1985) Riddles of natural categorization [and discussion]. Philos Trans R Soc Lond B Biol 308:129–144. CrossRef

Jensen G (2017) Serial learning. In: American Psychological Association handbook of comparative psychology, Vol 2 (Call J, Burghardt GM, Pepperberg IM, Snowdon CT, Zentall C, eds), pp 385–409. Washington, DC: American Psychological Association.

Jensen G, Altschul D (2015) Two perils of binary categorization: why the study of concepts can't afford true/false testing. Front Psychol 6:168. CrossRef Medline

Jensen G, Altschul D, Danly E, Terrace H (2013) Transfer of a serial representation between two distinct tasks by rhesus macaques. PLoS One 8:e70285. CrossRef Medline

Jensen G, Muñoz F, Alkan Y, Ferrera VP, Terrace HS (2015) Implicit value updating explains transitive inference performance: the betasort model. PLoS Comp Biol 11:e1004523. CrossRef Medline

Jensen G, Alkan Y, Muñoz F, Ferrera VP, Terrace HS (2017) Transitive inference in humans and rhesus macaques after massed training of the last two list items. J Comp Psychol. Advance online publication. Retrieved Mar 23, 2017. doi: 10.1037/com0000065. CrossRef Medline

Silverman JL, Gastrell PT, Karras MN, Solomon M, Crawley JN (2015) Cognitive abilities on transitive inference using a novel touchscreen technology for mice. Cereb Cortex 25:1133–1142. CrossRef Medline

Koscik TR, Tranel D (2012) The human ventromedial prefrontal cortex is critical for transitive inference. J Cogn Neurosci 24:1191–1204. CrossRef Medline

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, Vol 25 (Pereira F, Burges CJ, Bouton L, Weinberger KQ, eds), pp 1097–1105. Red Hook, NY: Curran.

Lazareva OF, Wasserman EA (2006) Effect of stimulus orderability and reinforcement history on transitive inference in pigeons. Behav Process 72:161–172. CrossRef Medline

Lazareva OF, Wasserman EA (2012) Transitive inference in pigeons: measuring the associative values of stimuli B and D. Behav Process 89:244–255. CrossRef Medline

Lazareva OF, Smirnova AA, Bagozkaja MS, Zorina ZA, Rayevsky VV, Wasserman EA (2004) Transitive responding in hooded crows requires linearly ordered stimuli. J Exp Anal Behav 82:1–19. CrossRef Medline

Lea SE, Ryan ME (1984) Feature analysis of pigeons' acquisition of concept discrimination. In: Quantitative analyses of behavior: discrimination processes (Commons ML, Herrnstein RJ, Wagner AR, eds), pp 239–253. Cambridge, MA: Ballinger.

Lee H, Grosse R (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proc Annu Int Conf Mach Learn 26:609–616.

Marsh HL, MacDonald SE (2008) The use of perceptual features in categorization by orangutans (*Pongo abelli*). Anim Cogn 11:569–585. CrossRef Medline

Matsukawa A, Inoue S, Jitsumori M (2004) Pigeon's recognition of cartoons: effects of fragmentation, scrambling, and deletion of elements. Behav Process 65:25–34. CrossRef Medline

McGonigle BO, Chalmers M (1977) Are monkeys logical? Nature 267:694–696. CrossRef Medline

McGonigle BO, Chalmers M (1992) Monkeys are rational! Q J Exp Psychol 45:189–228.

Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335:817–820. CrossRef Medline

Mur M, Ruff DA, Bodurka J, De Weerd P, Bandettini PA, Kriegeskorte N (2012) Categorical, yet graded-single-image activation profiles of human category-selective cortical regions. J Neurosci 32:8649–8662. CrossRef Medline

New J, Cosmides L, Tooby J (2007) Category-specific attention for animals reflects ancestral priorities, not expertise. Proc Natl Acad Sci U S A 104:16598–16603. CrossRef Medline

Newen A, Bartels A (2007) Animal minds and the possession of concepts. Philos Psychol 20:283–308. CrossRef

Roberts WA (1996) Stimulus generalization and hierarchical structure in categorization by animals. In: Stimulus class formation in humans and animals (Zentall TR, Smeets PM, eds), pp 35–54. Amsterdam: Elsevier.

Terrace HS (1984) Simultaneous chaining: the problem it poses for traditional chaining theory. In: Quantitative analyses of behavior: discrimination processes (Commons ML, Herrnstein RJ, Wagner AR, eds), pp 115–138. Cambridge, MA: Ballinger.

Terrace HS (2005) The simultaneous chain: a new approach to serial learning. Trends Cogn Sci 9:202–210. CrossRef Medline

Terrace HS, Son LK, Brannon EM (2003) Serial expertise of rhesus macaques. Psychol Sci 14:66–73. CrossRef Medline

Treichler FR, Van Tilburg D (1996) Concurrent conditional discrimination tests of transitive inference by macaque monkeys: list linking. J Exp Psychol Anim Behav Process 22:105–117. CrossRef Medline

Treichler FR, Raghanti MA, Van Tilburg DN (2007) Serial list linking by macaque monkeys (*Macaca mulatta*): list property limitations. J Comp Psychol 121:250–259. CrossRef Medline

Uchida N, Kepecs A, Mainen ZF (2006) Seeing at a glance, smelling in a whiff: rapid form of perceptual decision-making. Nat Rev Neurosci 7:485–491. CrossRef Medline

Van der Jeugd A, Goddyn H, Laeremans A, Arckens L, D'Hooge R, Verguts T (2009) Hippocampal involvement in the acquisition of relational associations, but not in the expression of a transitive inference task in mice. Behav Neurosci 123:109–114. CrossRef Medline

Van Opstal F, Verguts T, Orban GA, Fias W (2008) A hippocampal-parietal network for learning an ordered sequence. Neuroimage 40:333–341. CrossRef Medline

Vasconcelos M (2008) Transitive inference in non-human animals: an empirical and theoretical analysis. Behav Process 78:313–334. CrossRef Medline

Watanabe S (2013) Preference for and discrimination of paintings by mice. PLoS One 8:e65335. CrossRef Medline

Wendelken C, Bunge SA (2010) Transitive inference: distinct contributions of rostrolateral prefrontal cortex and the hippocampus. J Cogn Neurosci 22:837–847. CrossRef Medline

Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19:356–365. CrossRef Medline

Zalesak M, Heckers S (2009) The role of the hippocampus in transitive inference. Psychiatry Res 172:24–30. CrossRef Medline