PSYCHOLOGY

Reward associations do not explain transitive inference performance in monkeys

Greg Jensen[1,2]*, Yelda Alkan[2,3], Vincent P. Ferrera[2,3,4], Herbert S. Terrace[1]*

Most accounts of behavior in nonhuman animals assume that they make choices to maximize expected reward value. However, model-free reinforcement learning based on reward associations cannot account for choice behavior in transitive inference paradigms. We manipulated the amount of reward associated with each item of an ordered list, so that maximizing expected reward value was always in conflict with decision rules based on the implicit list order. Under such a schedule, model-free reinforcement algorithms cannot achieve high levels of accuracy, even after extensive training. Monkeys nevertheless learned to make correct rule-based choices. These results show that monkeys' performance in transitive inference paradigms is not driven solely by expected reward and that appropriate inferences are made despite discordant reward incentives. We show that their choices can be explained by an abstract, model-based representation of list order, and we provide a method for inferring the contents of such representations from observed data.

INTRODUCTION

According to Keynes, "Part of our knowledge we obtain direct; and part by argument" (*1*). Although few have put it as succinctly, he was hardly the first to observe that humans learn from experience as well as through logical inference. It is widely accepted that animals also learn from experience, but most studies of nonhuman animals assume that their choices depend chiefly on the expected reward value associated with choices (*2*–*4*). Rigorous demonstrations of "logical" learning in animals are rare (*5*, *6*) because reward associations are a nearly ubiquitous confound. Furthermore, despite growing evidence that nonhuman animals can manipulate representations by way of model-based learning (*7*–*9*), those representations are still derived from subjects' direct experience of the world, rather than from inferred relationships.

Transitivity is a property of ordered sets that, if exploited, can greatly reduce the amount of evidence needed to learn how any two items relate to one another. "Transitive inference" (TI) broadly refers to this nonassociative learning ability (*10*), and it has been displayed in every vertebrate species in which it has been tested (*11*). To demonstrate TI, subjects are first trained to choose between pairs of stimuli belonging to an ordered list (e.g., "ABCDEFG"). If subjects are trained only with adjacent pairs (e.g., AB, BC, etc.) but are subsequently able to judge the order of nonadjacent pairs (e.g., BD, CE, etc.), it appears as if they have acquired knowledge of the underlying list order as well as an understanding that the list order obeys transitivity. However, the interpretation of behavioral performance during TI tasks is controversial, and some theories posit that computations based on reward value play an important role. In the current study, we present a critical test of these hypotheses.

TI is a behavioral phenomenon whose underlying mechanism remains a topic of active debate. We use the term to refer to decision-making where preferences appear to rely on the transitive property of some ordering among the stimuli and cannot be explained by a reward-maximizing model-free learning process. Experiments using both human and nonhuman subjects have shown that performance on test pairs cannot be explained by expected reward value (*12*, *13*). Thus, the literature is unambiguous that behavior consistent with TI occurs in a wide range of species. What remains unclear is what cognitive mechanism might support this behavior. In addition, given the efficacy of model-free learning in many applied contexts, efforts to explain TI entirely in terms of reward associations have been remarkably persistent (*14*). An objective of the current study is to present subjects with an experimental procedure that cannot be solved by reward associations alone. By this, we mean that a model-free algorithm based on expected reward value should be unable to perform the task, regardless of the type or amount of training it receives.

We conjecture that the full scope of published TI results can only be explained if nonhuman animals perform inferences by manipulating an abstract representation of list order. Choices can be well approximated by treating each stimulus as having a position along a continuum and for the uncertainty of those positions to be described by probability distributions (*12*). This continuum is not derived from any specific properties of sensory experience, as we and others have used TI tasks that provide no spatial, temporal, or reward-based cues on which associations can be built. To demonstrate the efficacy of a model based on estimated position, we use it to describe TI performance when reward associations directly conflict with task performance, such that stimuli that are correct more often yield smaller rewards when chosen. Under these circumstances, a theory based on expected value would predict low overall accuracy.

We tested the ability of macaque monkeys ($N = 4$) to make inferences about the implied ordering of pairs of stimuli (Fig. 1, A and B). Subjects were always rewarded for choosing the stimulus with a lower implied rank. However, the amount of reward (water) was varied to distort expected value (Fig. 1C). In the "reverse reward gradient" condition, stimuli that were correct in more pairings yielded smaller rewards, such that in most pairings, the correct item had a lower overall expected value than the incorrect item. Expected value was a nonlinear function of list position, shaped like an inverted U. The conflict between the frequency with which a stimulus is correct and its reward size when it is correct should limit the efficacy of any strategy based on reward associations. In the "concordant reward gradient" condition, items that were correct more often yielded larger rewards, giving early items an exaggerated

[1]Department of Psychology, Columbia University, New York, NY 10027, USA. [2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA. [3]Department of Neuroscience, Columbia University, New York, NY 10027, USA. [4]Department of Psychology and Psychiatry, Columbia University, New York, NY 10027, USA.
*Corresponding author. Email: greg.guichard.jensen@gmail.com (G.J.); terrace@columbia.edu (H.S.T.)

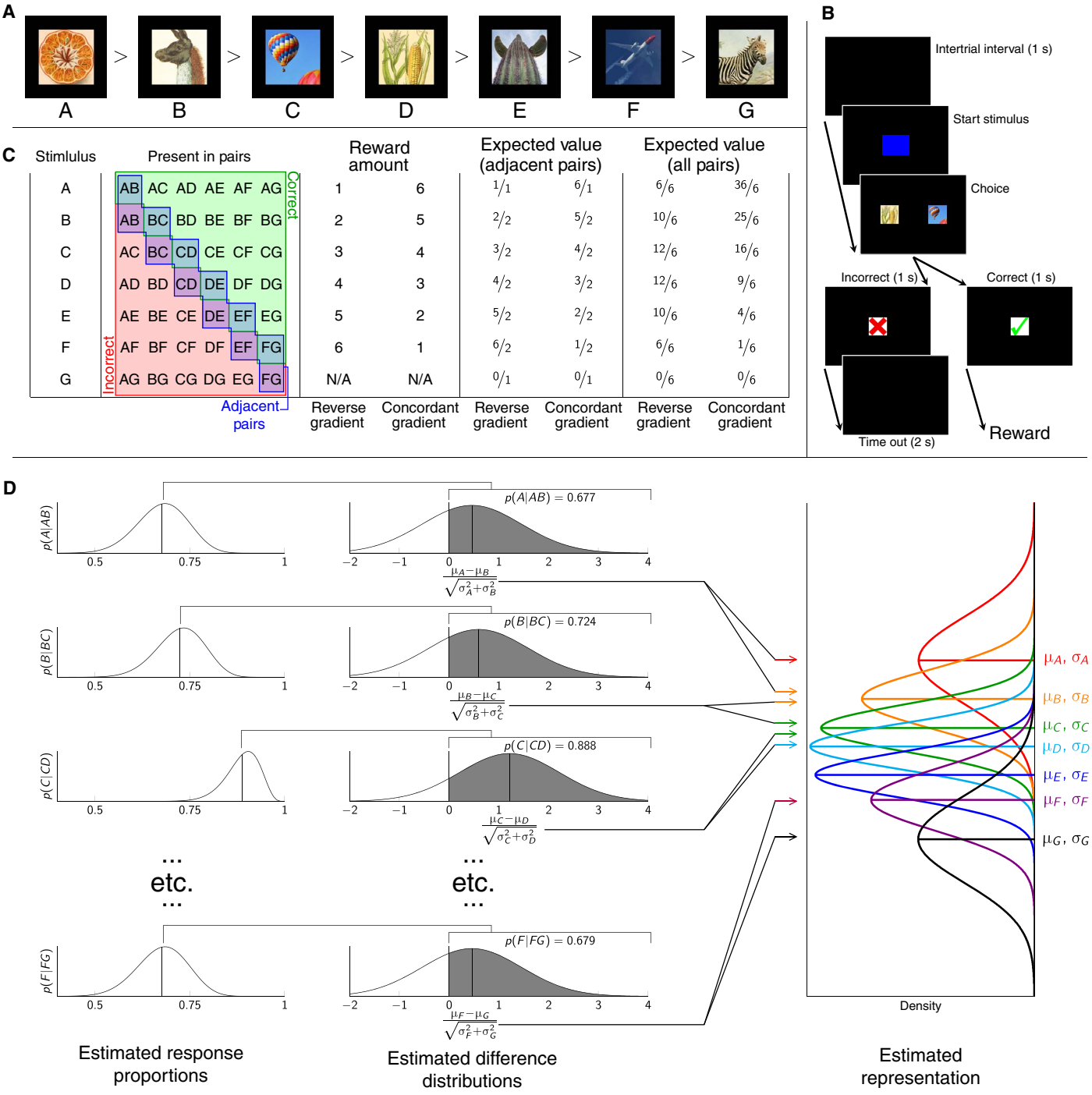

Fig. 1. Experimental and analytic procedure. (**A**) Subjects learned the ordering of seven-item lists, consisting of images (A, B, C, etc.). The correct item is always the item that occurs early in the list (e.g., B is correct for the pair BC). (**B**) Trial structure. After touching a start stimulus, subjects see two images. Touching the correct stimulus yields rewards of varying magnitude. Incorrect responses yield no reward and a brief time-out period. (**C**) Stimuli were presented in pairs. Training sessions presented only adjacent pairs, outlined in blue. Testing sessions presented all pairs. Reward amounts depended on the rank of the correct stimulus. The "reverse gradient" delivered one drop of water for correct responses to A, two drops for correct responses to B, etc. This gradient is labeled "reverse" because the overall expected value of F exceeds that of E, although choosing F when the EF pair is presented results in no reward. Thus, expected value cannot be used to guide which choice is correct. The "concordant gradient" delivered six drops for correct choices of A, five for B, etc. Therefore, the stimulus with the higher expected value is concordant with the correct choice. (**D**) Bayesian model for estimating stimulus position from observed response accuracy. Subjects are presumed to make use of a linear representation with uncertain stimulus positions. We assume that this representation takes the form of a normal distribution with some mean and SD for each stimulus. To infer these parameters, we estimated *p*(correct) for each pair and transformed this to the area above zero of some *z* distribution. Inferring the parameters in our representation is then done as a simultaneous estimation problem, implemented using Stan. Stimuli adapted from images in the public domain.

expected value. Across all pairs, this produced a scallop-shaped expected value function. Since reward is a useful cue in this condition, high levels of accuracy are expected.

To consistently solve the task under both conditions, subjects had to learn the list position of each item while simultaneously disregarding its expected value. We have previously shown that such learning can be supported by representing the position of each stimulus along a continuum (*12*). Using the patterns of response accuracy among the stimulus pairs, one may infer the positions and uncertainty of each stimulus along this hypothetical spatial continuum that best recreate those errors (Fig. 1D). This was accomplished by implementing a model based on an internal representation of item position, solved as a simultaneous estimation problem using the Stan language (*15*). Whatever computation is being performed to achieve both the success rates and patterns of response error in this and other published tasks needs to behave as though feedback from each trial provided evidence about list position, rather than reward value. No model-free learning algorithm gives a good description of this behavior. By rendering our cognitive model in computationally rigorous terms, we are able to make specific predictions about performance.

RESULTS

Contrary to a prediction based on expected reward value, subjects consistently learned the list orderings (Fig. 2A). In both conditions, performance exceeded chance accuracy by the end of training and remained above chance at the start of all-pairs testing. Although the reverse gradient appeared to slow the learning rate, performance nevertheless reliably rose to above-chance levels by the end of training. This high performance is especially noteworthy under reverse gradients, because it meant that subjects consistently chose stimuli that were worth a smaller amount overall, as judged by their own learning history.

To evaluate how decisions based on reward value alone would be made, two model-free *Q*-learning algorithms (*12*, *16*) also performed the task, each using a softmax decision rule (*17*, *18*). One algorithm (plotted in red) used the parameters that provided the best fit to the

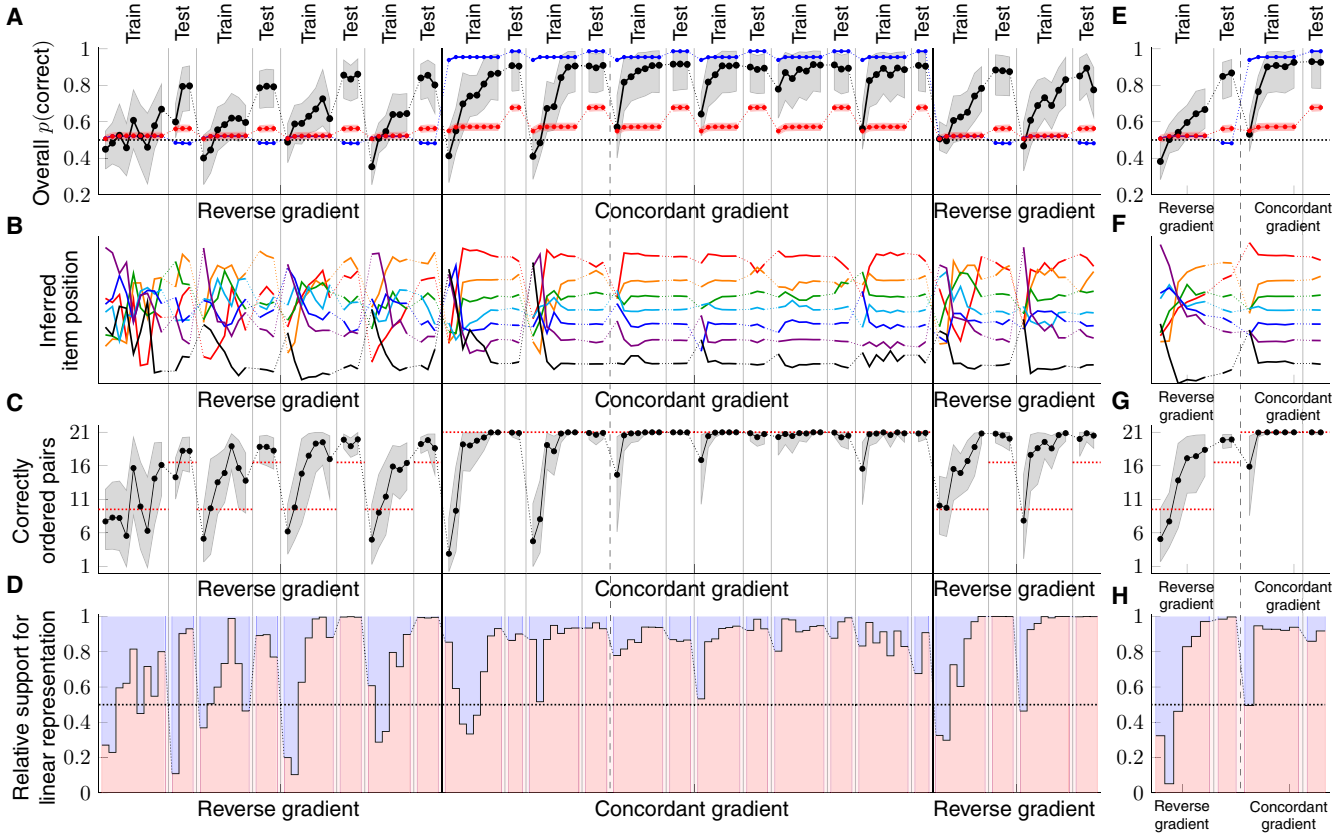

Fig. 2. Subjects retained the same list of stimuli for each set of training and testing sessions and then were presented with a new list at the start of the next training session. Vertical dashed lines indicate a break lasting 1 year. (**A**) Population estimates of response accuracy (black) for each session. Chance is indicated by the dotted lines. Red circles correspond to the exploratory *Q*-learner, fit to the observed data. Blue circles correspond to the exploitative *Q*-learner, based on a previous study (*12*). (**B**) Peaks of inferred position distributions of each stimulus in subjects' representations. Red, A; orange, B; green, C; cyan, D; blue, E; violet, F; and black, G. Subjects reconstructed the stimulus order in the concordant gradient condition, and did so approximately in the reverse gradient condition, with the exception of stimulus A. (**C**) Average number of the 21 stimulus pairs that fell in the correct order, based on the model estimates in the above panel. The red dotted line indicates how many pairs would be ordered correctly if subjects used expected values as the basis for ordering. (**D**) Support of the evidence for the positions in the inferred representation being organized according to a strictly linear representation (in red), relative to the expected values (in blue), according to a Bayesian Information Criterion (BIC) analysis. Subjects tended to be equivocal, or to favor the expected value, during the first few sessions of training. However, late in training, and throughout the testing phase, the inferred representations more closely resembled a linear ordering of stimuli with uniform spacing. (**E** to **H**) Same as (A) to (D), respectively, but based on pooling data across the six lists.

observed data. Its learning rate was slow, and its decision rule favored exploratory behavior. This enabled it to exceed chance by making productive mistakes when the reverse gradient was in effect. However, the algorithm's high error rate also prevented it from achieving the high success rate displayed by subjects in either of our experimental conditions.

The second algorithm (plotted in blue) used parameters that maximized total rewards. Its learning rate was rapid and its decision rule favored winner-take-all exploitation of the largest expected value. This yielded near-perfect performance in the concordant gradient condition but led to performance slightly below chance in the reverse gradient condition. Subjects outperformed both algorithms overall. Critically, neither extreme (nor any intermediate parameter values) allowed *Q*-learning to do better than 60% accuracy in the reverse gradient condition.

It is important to note that the *Q*-learning algorithms are unable to solve the TI task throughout the testing phase of the reverse gradient condition, not just at its onset. In traditional TI experiments that use uniform rewards, a model-free learning algorithm that is being presented with all pairs quickly learns the order of the stimuli based on their expected value, even if performance on critical pairs was at chance levels at the end of training (*12*). This is why most studies of TI focus on accuracy at the moment of transfer, not throughout a prolonged testing phase. However, the reverse gradient condition cripples *Q*-learning throughout testing, keeping it below 60% accuracy even after almost 2000 trials presenting all pairs. The consistently high accuracy of subjects in this condition requires some other explanation than using experienced reward as a proxy for list order.

An alternative to *Q*-learning is a model-based algorithm that represents the list position rather than the reward value of each item. It is important to note that for both reward conditions (reverse and concordant), average reward magnitude (Fig. 1C) was a nonlinear function of item position. Therefore, in both conditions, the effect of expected value was dissociable from the effect of position. We were able to infer each item's position within the subject's representation by examining the particular patterns of error among all pairs of stimuli experienced in a session. As shown in Fig. 2 (B and F), the implied stimulus positions suggest that expected value did have an influence in the early stages of training, especially in the reverse gradient condition. Over the course of training, however, subjects gradually worked out the approximate order. Consequently, subjects appeared to rely on inferences about stimulus position, rather than expected value.

To further demonstrate that subjects behaved as though they represented item positions that were uniformly spaced along a continuum, we used the inferred content of the representations to calculate the average number of pairs expected to yield correct responses (Fig. 2, C and G). Since this prediction is based on our theoretical model of item position, we can estimate implied performance for all 21 pairs even during training sessions when only six pairs were presented. In the reverse gradient condition, subjects were expected to get more pairs right than would be predicted by comparing their expected values (as shown in Fig. 1C).

We also used Schwartz-Bayes information criterion scores (*19*) to measure the relative support of the evidence for whether the contents of the representation (in Fig. 2, B and F) were better estimated by the expected values (as in Fig. 1C) or by a uniformly spaced linear model of list position (as in Fig. 1D). Early sessions of training tended to favor the expected value distributions, but the linear representation better accounted for the patterns of behavior by the second session (for concordant gradients) or third session (for reverse gradients) of training. The linear representation was then consistently favored during testing.

Although Fig. 2 supports the hypothesis that subjects behaved as though stimuli were arranged into an approximate order, it provides few details about the particulars of behavior itself. It also does not give a comprehensive picture of how behavior manifested. In broad terms, there are three questions that arise naturally in the present study: (i) What is the overall probability of a correct response, (ii) to what extent did the reward gradient influence response accuracy, and (iii) to what extent was response accuracy predicted by the difference between stimulus ranks? This last measure, formally studied as the "symbolic distance" (*11*, *20*), is important because it has been widely reported that accuracy tends to be higher for stimuli that are separated by larger gaps. Since we expect, on the basis of the literature, for performance to be positively correlated with symbolic distance, it needs to be measured and controlled for.

Session-by-session estimates of response accuracy (Fig. 3A) were obtained using binomial regression, as described in Materials and Methods. These estimates show that, across stimulus pairs, response accuracy tended to grow over the course of training (as expected), with comparatively high performance during testing sessions. The population mean effects of reward magnitude were estimated for all (training and testing) sessions, and the effects of symbolic distance were estimated for testing sessions (Fig. 3B). Although these effects were very uncertain (due to being limited to one session's worth of data per estimate), the estimated size of reward effects was smaller than that of symbolic distance during almost every testing session.

To get a more precise estimate of these effects, binomial regressions were performed for each step in a training cycle, pooled across all phases under a particular condition. These show more clearly that response accuracy tended to grow more slowly in the reverse gradient condition than in the concordant gradient condition (Fig. 4A), but that overall response accuracy was well above chance at test in both conditions. In addition, although distance effects were reliably observed in both conditions (Fig. 4B), the effect size of the reward gradient effect was small throughout training (Fig. 4C) and was especially uncertain during testing phases. Mean subject-level estimates are also provided (Fig. 4, D and F).

The contrast in performance in the two conditions suggests that the reverse gradient condition interfered with performance relative to the concordant gradient condition, as evidenced by the higher intercept estimates in the latter case. To assess the consistency of this effect, we calculated parameter differences for population-level (Fig. 4, G to I) and subject-level (Fig. 4, J to L) parameters, with corresponding uncertainty. In most cases, the 95% credible intervals for the differences overlapped with zero, but a notable exception was the intercept term during both phases of testing. The consistently positive values for the intercept differences indicate consistently higher response accuracy in the concordant gradient condition than in the reverse gradient condition.

To see the implications of the models described in Fig. 4, we compared the estimated response accuracy of each stimulus pair (with corresponding uncertainty) to the observed empirical mean response accuracy for the last session of training and the first session of testing (Fig. 5). Plotted in this way, it is clear that, in general, the model expects that the reward manipulation imposed a tilt to the response accuracies in the reverse gradient condition, which is visible both in the empirical rates and in the model estimates. However, the concordant gradient results are much more equivocal about whether a

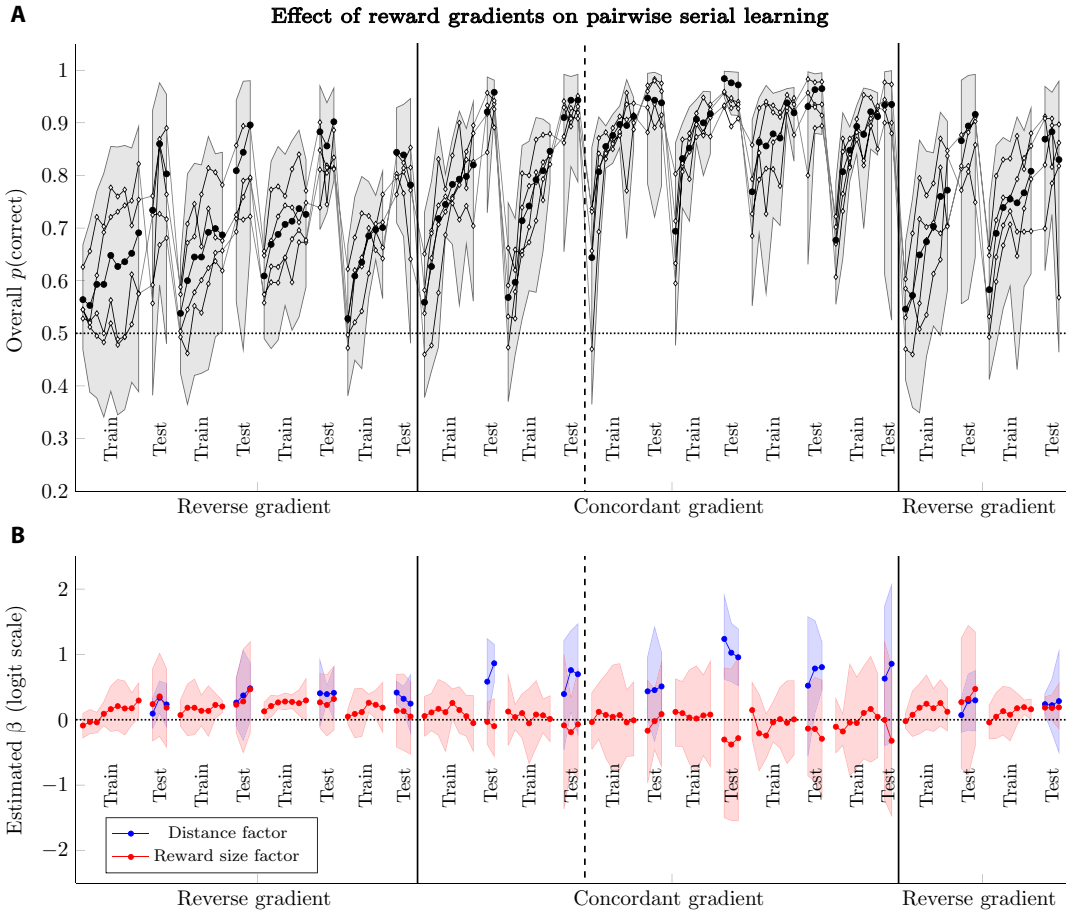

Fig. 3. Results of session-by-session binomial regression estimation of response accuracy. Shaded intervals represent the 95% credible interval for the population estimate. (**A**) Overall estimated response accuracy for each session (black circles), as well as subject-level estimates of performance (white diamonds). (**B**) Estimated regression parameters for the effects of symbolic distance (blue) and reward magnitude (red), plotted on the logit scale. Distance effects could only be estimated for testing phases, as all stimulus pairs during training had the same symbolic distance.

reward effect is evident in the population. If anything, the empirical estimates suggest a very mild inverted U shape, although the uncertainties are large enough that this effect may merely be a product of experimental noise.

The results above are suggestive that subjects successfully made a TI in their transition from training to testing, but those estimates of performance are overall for each session. To test whether subjects were above chance on the first trial of testing (before further learning could take place), it is necessary to work with a much more constrained dataset. It is also important to focus on the "critical test" pairs (i.e., those that do not include terminal items and were not part of the training set). For the first presentation of each of the six critical pairs, subjects in the reverse gradient displayed an effect that was close to chance (Fig. 6A), while those in the concordant gradient condition showed a clear preference for the correct answer (Fig. 6D). However, because the number of cases was so small (six trials per subject per estimate), the uncertainties surrounding this estimate are very large. To better leverage the available data, we used logistic regression models (described in Materials and Methods) that could track the learning rate and thereby project the response accuracy on the first trial of testing to be extrapolated. When such models were fit to each pair separately, the results were suggestive of responding above chance in the reverse gradient condition (Fig. 6B) and clearly above chance in the concordant gradient condition (Fig. 6E). In addition, a "full regression model" fit the accuracies of all critical pairs in both conditions, taking symbolic distance and reward size into account. These estimates (Fig. 6, C and F) also suggest that subjects made correct responses to these critical pairs at the start of the testing sessions.

DISCUSSION

This is the first study to test reward-based explanations of TI performance in a paradigm in which subjects have to ignore the experienced magnitude of reward to choose the correct response. Subjects' success in learning under the reverse reward gradient adds to a growing body of demonstrations that nonhuman animals can solve inference-based problems by a more cognitive means than favoring whichever alternative has the highest experienced value. Other results consistent with this kind of abstract serial inference include robust learning when some pairs are presented more often than others (*13*, *21*), combining separate lists by training a single linking pair (*22*), and transfer of serial knowledge from one experimental paradigm to another (*23*, *24*). These studies have usually emphasized inference at the moment of testing (when the nonadjacent pairs are entirely novel), because although model-free algorithms have difficulty during training, most are able to solve the task over the subsequent trials of the testing

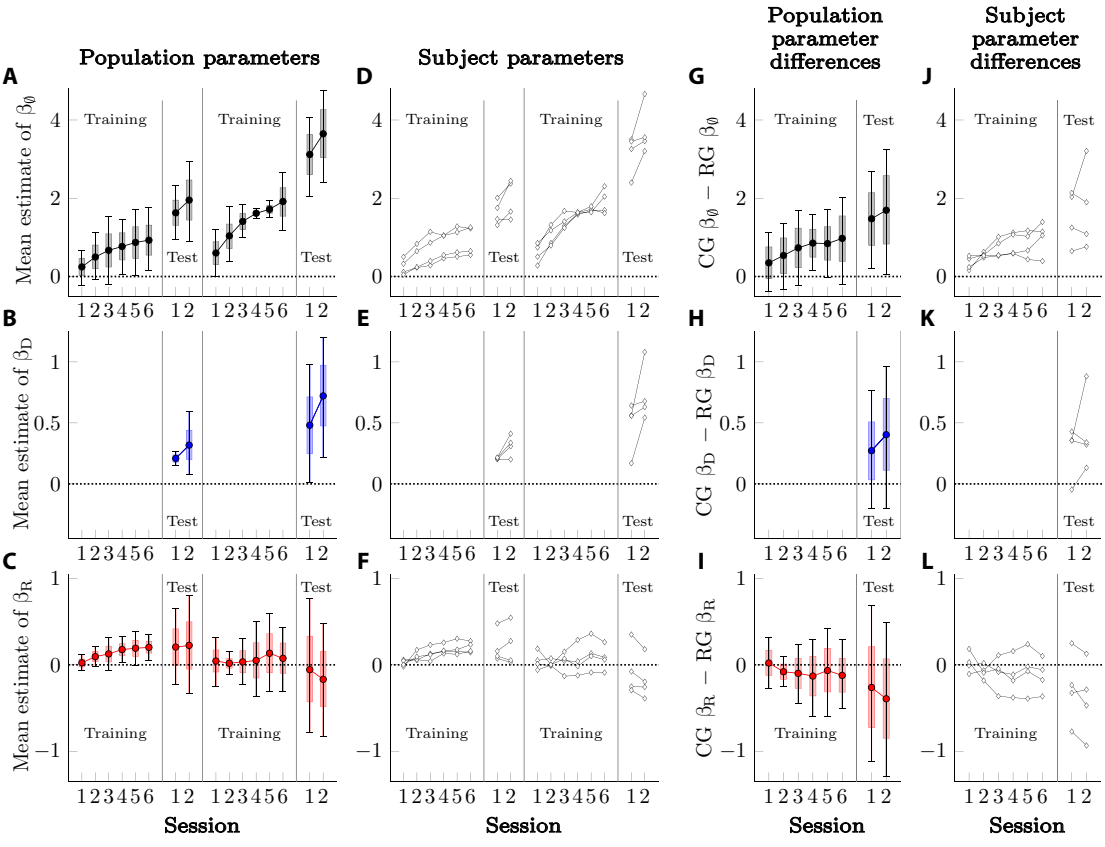

Fig. 4. Parameter estimates for binomial regression of response accuracy, pooled across the six lists learned in a given phase, plotted on the logit scale. Boxes represent the 80% credible interval for the population estimate, while whiskers represent the 95% credible interval. (**A**) Population estimates of the session intercepts, representing overall response accuracy (i.e., a value of 0.0 denotes chance performance). (**B**) Population estimates of the effect of symbolic distance. (**C**) Population estimates of the effect of reward magnitude. (**D** to **F**) Same as (A) to (C), respectively, but subject-level parameters estimated for each of the four subjects. (**G** to **I**) Posterior differences between the population-level parameters of the concordant gradient and reverse gradient conditions. (**J** to **L**) Posterior differences between the subject-level parameters.

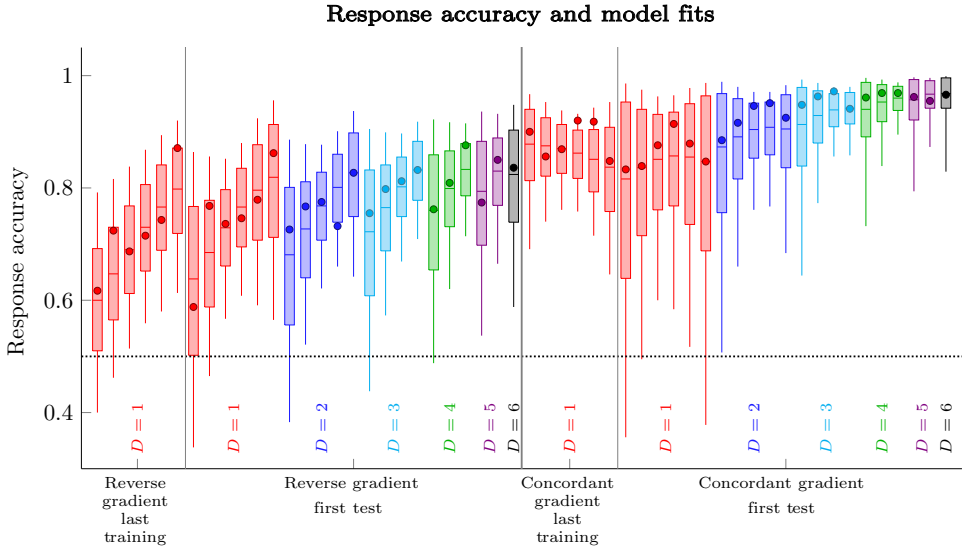

Fig. 5. Response accuracy for all pairs presented during the last session of training and the first session of testing, based on the parameters reported in Fig. 4. Circles represent the raw empirical estimates, independent of the model fit. Boxes are centered at the mean population accuracy, with their upper and lower extent representing the 80% credible interval for the population estimate. Whiskers represent the 95% credible interval. Pairs are first organized by symbolic distance (red, distance 1; blue, distance 2; etc.) and then alphabetically (AB, BC, CD, etc.).

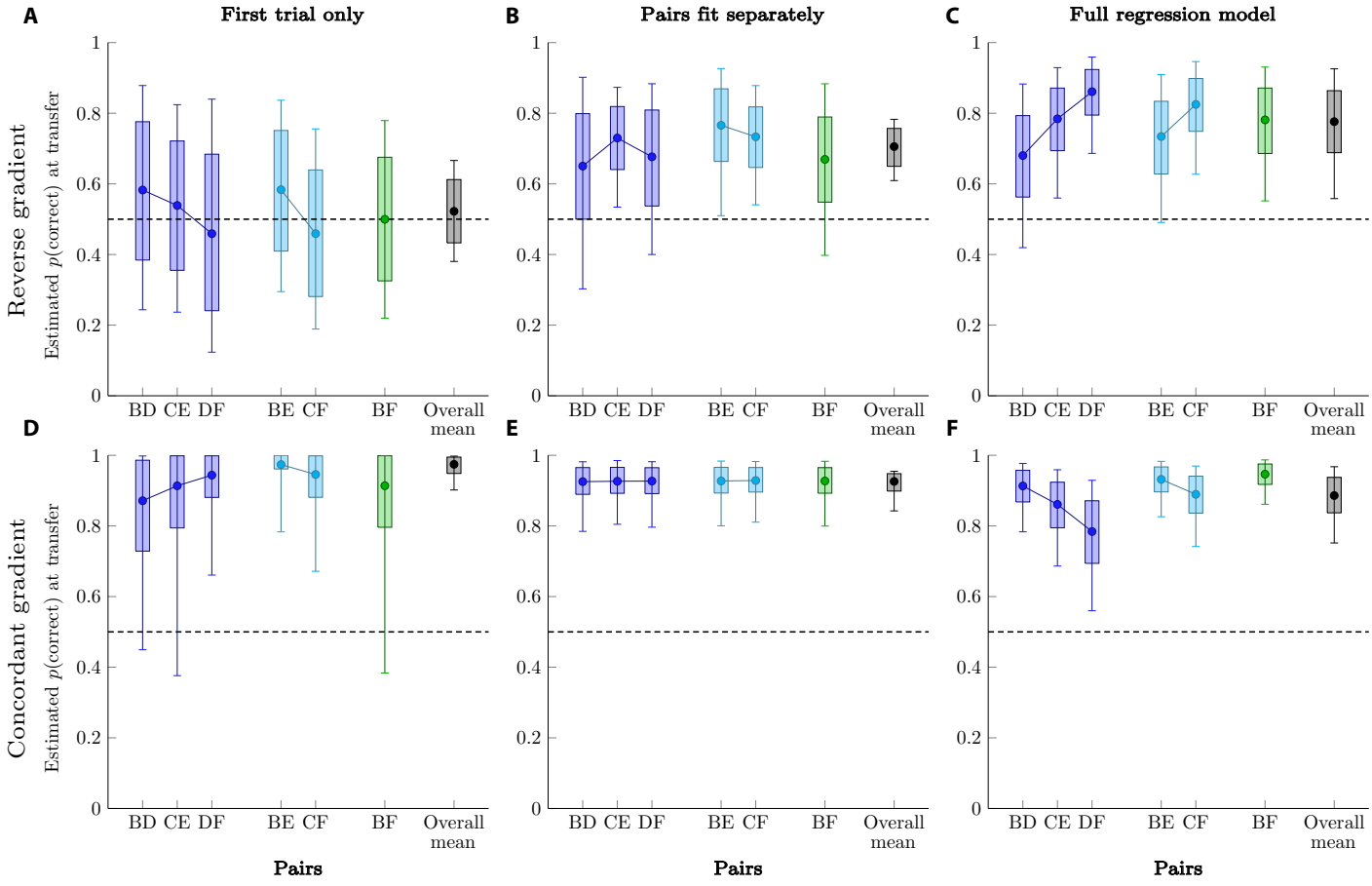

Fig. 6. Estimated response accuracy for critical pairs at the trial of testing. Boxes are centered at the mean population accuracy, with their upper and lower extent representing the 80% credible interval for the population estimate. Whiskers represent the 95% credible interval. (**A**) Population estimates of critical pair accuracy in the reverse gradient condition based on a binomial regression using only the first presentation of each pair. In addition, mean accuracy across all critical pairs, with corresponding uncertainty, is plotted in black. (**B**) Population estimates in the reverse gradient condition based on logistic regressions, extrapolating performance at the intercept (trial zero). Each pair was fit separately in this analysis, and the mean of the critical pairs was pooled across those estimates. (**C**) Population estimates of trial zero performance in the reverse gradient condition based on a logistic regression that incorporated data from all six pairs in both conditions, fitting both distance and reward effects. (**D** to **F**). Same as (A) to (C), respectively, but associated with the concordant gradient condition.

phase. The reverse gradient is an exception: Model-free learning is never able to fully solve the task. Over sustained testing of all pairs, *Q*-learning can do no better than either 60% accuracy (when highly exploratory) or responds below chance (when highly exploitative). Although model-free algorithms can improve that flexibility in many scenarios by adjusting their exploration/exploitation balance, this flexibility is not sufficient to achieve performance comparable to that of our subjects in the reverse gradient condition.

Although the present study uses a TI procedure, it has broader implications for trial-and-error learning in animals. The standard test of TI in animals, in which adjacent pairs of a five-item list are trained and critical pairs are subsequently tested [e.g., (*25*)], was designed with the express purpose of equalizing experienced rewards for all nonterminal items. This makes performance at the start of the testing phase (before a differential reward effect) the focus of the study—if subjects exceed chance on the critical pair BD, this demonstrates a reliance on something other than expected value. Subsequent studies that have used longer lists have identified distance effects at transfer among the critical pairs [e.g., (*12*)], but those studies remain focused on the effect at transfer because that was the moment that an expected value strategy was expected to yield chance performance. In the present experiment, however, high levels of performance at any stage of reverse gradient training or testing should not have been possible for a reward-driven algorithm, as demonstrated by our *Q*-learning algorithms. Even if we had presented all pairs from the outset of training, with no critical test of TI included, an algorithm that chooses on the basis of stimulus-reward associations would still perform poorly, far below the levels observed in our subjects. As such, although we find evidence of TI in the present data (Fig. 6), this is merely part of a wider pattern of behavior inconsistent with reward-maximizing associative behavior throughout the reverse gradient condition.

Just as nonhuman subjects appear to rely on an "approximate number system" (*26*) to compare quantities, rather than perform fully abstract arithmetic, the literature on serial learning suggests that animals are also equipped with an "approximate reasoning system" that is able to make simple logical inferences, even when reward signals provide conflicting information. While we do not argue that nonhuman animals are engaging in formalized propositional logic, their performance can nevertheless be explained only by a model-based system. We recognize that although our spatial model is one candidate

model, other cognitive mechanisms could, in principle, be responsible for the present behavior. The present data alone cannot, for example, rule out that subjects learned ad hoc rules to favor large rewards in some cases and small rewards in others, changing their behavior as circumstances demanded. We find this account unlikely, for two reasons. First, if subjects depend on ad hoc rules, we would expect more individual variation in performance; instead, performance was very consistent across subjects (Figs. 3A and 4, D to F). Second, performance in this study was qualitatively similar to that seen in the extensive literature on TI in animals (*10*, *11*). This suggests to us that the mechanism underlying the present performance is not different in kind from that seen in other studies, most of which did not manipulate reward magnitude.

With this in mind, it is important to consider the implications of the current results. What they cannot support is a claim that the high levels of accuracy achieved in the reverse gradient condition (which display both learning rates and distance effects seen in other studies of TI) can arise solely from model-free learning based on maximizing expected value, the mechanism at the heart of *Q*-learning and other associative models. Instead, our results are best explained by the joint influence of the two systems, one reward-driven and one model-driven, acting in parallel. Advances in our understanding of brain circuitry has helped us to move past the either/or logic of the cognitive revolution [e.g., (*27*)]. Furthermore, the present results suggest that performance is likely not merely the passive averaging of these two systems. It appears as though associative mechanisms interact with inference in some cases, such as the distortion of pairwise accuracy we observed in the reverse gradient condition (Fig. 5), whereas those associations are at other times ignored entirely, as in the case in the concordant gradient condition, during which performance better approximated a linear arrangement of stimulus positions than a scalloped distribution of expected values.

Similarly, associative models predict that when some stimulus pairs are presented much more often than others, a substantial distortion in subsequent TI should be observed. However, despite evidence that such associative signals are likely calculated, studies that included massed presentations of certain stimulus pairs yielded no evidence at all of distorted inference (*11*, *13*). While it should no longer be controversial that subjects make both cognitive inferences and expected value calculations, two questions remain: How do these two calculations interact and when does one supersede the other in determining behavior?

Our approach is also distinct from previous studies of TI because our computational model is not limited to qualitative statements about whether subjects can perform TI. By mapping a link between estimates of behavior and a formal model of the representation's contents (Fig. 1D), it is possible to probe the consequences of experimental manipulations in more detail than ever before. In general, cognitive models in the TI literature do not make specific enough predictions about behavior to weigh them fairly against other alternatives (*24*). Our approach is the exception to that rule. It is our hope that it will not only help put to rest the long-standing objection that TI in nonhuman animals must be explainable by some reward association mechanism but also illustrate how a fully realized computational model can explain a cognitive phenomenon like TI in animals.

MATERIALS AND METHODS

Subjects

Subjects were four adult male rhesus macaques, N, O, R, and S. All subjects had prior experience with serial learning procedures, including TI. However, subjects had not previously been exposed to manipulations of reward magnitude in the context of serial learning.

Subjects were housed individually in a colony room, along with approximately two dozen other macaques. Experiments were performed in their home cages, using the apparatus described below. To increase motivation, subjects were fluid-restricted to 300 ml of water per day, or however much they were able to obtain by performing the task, whichever was greater. Typical performance yielded between 200 and 300 ml, whereas perfect performance could yield as much as 500 ml. As needed, supplemental water was given to subjects each day after the end of the experimental session. Monkeys were also given a ration of biscuits (provided before experimentation each day) and fruit (provided after experimentation).

The study was carried out in accordance with the guidelines provided by the *Guide for the Care and Use of Laboratory Animals* of the National Institutes of Health (NIH). This work, carried out at the Nonhuman Primate Facility of the New York State Psychiatric Institute (NYSPI), was overseen by NYSPI's Department of Comparative Medicine and was approved by the Institutional Animal Care and Use Committees (IACUCs) at Columbia University and NYSPI.

Apparatus

Subjects performed the task using a tablet computer. The tablet, running Windows 8.1, presented subjects with a 10.1″ HD display (1266 × 768 resolution), which both presented stimuli and provided a capacitive touch screen interface to record responses. All tasks were programmed in JavaScript and run using the Google Chrome browser, set in kiosk mode.

To deliver rewards, the tablet was connected to a solenoid valve by way of an Arduino Nano interface, which opened the valve for fixed intervals when rewards were delivered via a steel spigot below the tablet. One "drop" of water corresponded to 0.25 ml of fluid. When subjects received multiple drops, the valve opened and closed that many times in rapid succession to ensure that a consistent volume of liquid was being delivered. This apparatus was mounted in a Lexan frame, which fit securely into the space created by opening the door to the subject's home cage. Unless otherwise noted, this device was identical to that described in previously published studies (*28*).

At the start of each trial, a solid blue square was presented in the center of the screen, in order to focus the subject's attention and to direct their hand toward a consistent center point. Touching it initiated the next trial. All experimental stimuli were 250 × 250 pixel images, presented to the right and left of the start stimulus (Fig. 1B).

Procedure

Stimulus lists were assembled, consisting of fixed sets of seven photographic images apiece (list orders are hereafter identified as ABCDEFG, e.g., Fig. 1A). During each trial, two stimuli were presented simultaneously. The stimulus whose rank came earlier in the ordered list would always, if selected, yield a reward. The stimulus whose rank came later never yielded a reward. Subjects had to learn the ordering of the images by trial and error.

Each session consisted of up to 600 trials (fewer if the subject stopped responding before finishing the session). The experiment was divided into training phases (during which only adjacent pairs AB, BC, CD, DE, and FG were presented) and test phases (during which all 21 pairs were presented). During training, each session was organized into "blocks" of 12 trials each. During a block, each of the six adjacent pairings (listed above) were presented twice, once for each spatial arrangement (e.g., for

the pair AB, every trial in which A was on the left and B was on the right was balanced by another trial with B on the left and A on the right). These two trials might be presented in either order and with any number of intervening trials. Spatial counterbalancing was done for every pair of items. During testing, each session was organized into blocks of 42 trials each, counterbalancing the on-screen arrangement of the 21 possible pairs in a similar fashion. Subjects completed one session of the experiment a day and in almost all cases completed all 600 trials.

Sessions consisted of one of two reward conditions. The first was the "reverse gradient" condition, in which a correct response earned the number of drops equal to its rank (Fig. 1C). In all cases, an incorrect response yielded no reward. Thus, for example, when presented with the pair AB, a response to A earned one drop (because A has a rank of 1), whereas a response to B would earn no drops (because it is incorrect). However, when presented with the pair FG, a response to F would earn six drops (because its rank is 6), and a response to G would earn no drops. As a result, the reward value for responding correctly to each pair varied directly with the rank of that pair.

The second condition was the "concordant gradient" condition, in which stimuli of a lower rank yielded larger rewards (Fig. 1C). Such a gradient is "concordant" in that the reward amount associated with a stimulus is positively correlated with the odds of that stimulus being correct in a random pairing. Correct responses to A yielded six-drop rewards, correct responses to B yielded five-drop rewards, and so on, until correct responses to F yielded one-drop rewards.

At the start of each training phase, a new list of seven unfamiliar stimuli was used. Between five and nine sessions of training were then followed with two to three sessions of testing. This was collectively considered a "training cycle." Subjects learned a total of 12 lists over the course of the experiment. Advancing from a training phase to a testing phase was based on constraints of the academic calendar and not on measures of performance. As such, the start of testing was not contingent on a "learning criterion."

Subjects first learned four lists under the reverse gradient condition (each having a training phase and a testing phase). This was followed by two lists learned under the concordant gradient condition. After these six lists were completed, subjects took a 1-year break from the experiment (during which they participated in experiments that had no differential gradients of reward magnitude). Following this break, they learned four lists using the concordant gradient, followed by two lists using the reverse gradient. The purpose of the break was to provide a control against order of learning effects, since it was unclear whether a reverse-to-concordant transition would result in similar performance, compared to a concordant-to-reverse transition.

Statistical analysis: Bayesian model of stimulus position

To provide a computationally tractable Bayesian model of behavior, it was assumed that the position of each stimulus was represented by a normal distribution with parameters μ_i and σ_i for stimulus *i*. On each trial, a random value was drawn from each distribution, and the stimulus that was larger was selected. When distributions overlapped, the distribution with the higher mean was more likely to be selected, but the alternative items were still chosen some amount of the time. This recapitulates the logic behind the betasort model (*12*) but made use of normal distributions instead of beta distributions to facilitate parameter estimation.

Like the betasort model, there was also a parameter specifying the probability that subjects would disregard the representation and make a completely random response. This was included because monkeys in many experimental contexts never achieve perfect performance, instead maintaining some error rate regardless of how much additional training they receive. This parameter is denoted by θ, where $0.0 < \theta < 1.0$.

The odds that one normal distribution yields a larger value than a second normal distribution are identical to the odds that the difference between the two values is positive. Since the variances of normal distributions are additive under subtraction, and since there was a probability of θ of an arbitrary response, the overall probability of a stimulus A in the pairing AB is given as follows

$$p(A \mid AB) = \frac{\theta}{2} + (1-\theta)\int_{0.0}^{\infty} \mathcal{N}\left(x \mid \mu_A - \mu_B, \sqrt{\sigma_A^2 + \sigma_B^2}\right) dx \quad (1)$$

Since there are seven stimuli, behavior is modeled in terms of 15 parameters: μ_i and σ_i for each stimulus *i*, as well as θ. Estimating these requires solving a simultaneous estimation problem, where every pair of stimuli has its own version of Eq. 1. Figure 1D depicts how this simultaneous estimation can translate the observed response accuracies for all pairs to the corresponding estimates of the position of each stimulus. The electronic supplement includes a model coded using the Stan programming language (*15*) that solves this problem as a multilevel model, yielding estimates of these parameters for both the population and for each subject.

Note that because μ_i is unitless and defined only relative to the means of the other stimuli, values of μ_i and σ_i may be rescaled arbitrarily, as long as the scale is applied consistently for all position parameters. To give a common scale to the positions for the purposes of plotting their estimates, the μ_i parameters were centered at zero and then all μ_i and σ_i parameters were divided by the sample SD of the μ parameters. From a performance perspective, the model is unchanged under rescaling, since its comparisons are all relative between stimuli. The parameter estimates given in the results above (Fig. 2, B and F) are the population means, rescaled in this way. Figure 2 presents two additional ways of validating that performance of the Bayesian model is not consistent with expected reward value: a count of the "number of correctly ordered pairs" and an information criterion measure.

The estimate of the number of correctly ordered pairs depends on three details. First, the expected values of the stimuli during training were not the same as during testing, so the number of pairs an expected reward value comparison would order correctly changed. Second, despite there being only six training pairs, the model is always capable of making inferences about all 21 pairs. Consequently, the estimates give an idea of how many of the 21 pairs would be correct if the next trial was the start of testing. Third, because the parameters estimated using Stan, they took the form of chains of Markov Chain Monte Carlo (MCMC) results, not parameterized distributions. Since these chains of estimated value capture the covariation of estimates, it is valuable to assess how many pairs are correctly estimated for each iteration of the chain and to use the distribution of resulting values to obtain a credible interval for the estimate. In the event of ties in expected value (e.g., A = $^1/_1$ versus B = $^2/_2$), the pair is given a value of 0.5 since it is expected to be chosen correctly half the time. Although, on average, the nominal expected reward value comparisons identify the correct order of most pairs during the testing phase of the reverse gradient condition, subjects outperform even that higher bar (Fig. 2, C and G).

To determine whether the "weight of the evidence" better favored a uniformly distributed linear representation or one based on the expected reward value, linear models were fit using the nominal values

of each model as predictors and the posterior distributions of stimulus positions from the Bayesian model as the outcomes (linear on the one hand or as expected from the expected value on the other). Each regression yielded a BIC score. The relative support of the evidence for the linear model over the expected value model, on a scale of 0.0 to 1.0, is given as follows

$$\text{Support for Linear Representation} = \frac{\exp(-BIC_{\text{linear}})}{\exp(-BIC_{\text{linear}}) + \exp(-BIC_{\text{expected value}})} \quad (2)$$

Since the output of the MCMC analysis was a chain of position estimates, BIC scores were calculated for each iteration of the chain. Figure 2 (D and H) reports the mean BIC scores, averaged across all values in the chain.

Statistical analysis: Q-learning simulation of reward-driven behavior

Rather than merely assert rhetorically that expected or experienced reward value is insufficient to solve the reverse gradient condition, we used a model-free *Q*-learning algorithm (*16*) that can only use its experienced estimate of each item's reward value to perform the task. Although *Q*-learning ordinarily factors the "maximum possible reward on the next trial" into its updating, TI tasks are scheduled in such a way that a subject's choice on trial t has no impact on which choice alternatives are available on trial $t + 1$. Consequently, this "projection into the future" cancels itself out, leaving only basic reward prediction error updating of the memory vector Q

$$Q(\text{choice}) = (1 - \alpha)Q(\text{choice}) + \alpha \cdot \text{Reward} \quad (3)$$

That is, on each trial, the algorithm uses the reward delivery (including a value of 0.0 when no reward is delivered) to update its value of the item chosen. Items that are not chosen are not updated.

Equation 3 describes the memory vector, but not the criterion by which choices are made. In our implementation, choices are made on the basis of the softmax function (*17*, *18*), with an exponential term β

$$p(A \mid AB) = \frac{\exp(Q(A)^{\beta})}{\exp(Q(A)^{\beta}) + \exp(Q(B)^{\beta})} \quad (4)$$

When $\beta < 1.0$, the result is an algorithm whose behavior is more exploratory, because it is more willing to choose response option B when A has a larger expected value. At the extreme, when $\beta = 0.0$, subjects are equally likely to choose A and B. Contrastingly, if $\beta > 1.0$, the algorithm will behave in an exploitative manner, because it will be biased more strongly toward the larger expected value. As β tends toward large values, preference strongly tilts toward exclusive selection of the item with the highest expected reward value.

Using the Stan programming language (*15*), performance of each subject was fit using Eqs. 3 and 4. The best-fitting parameters were α values of between 0.086 and 0.225, while the best-fitting β parameters values were between 0.365 and 0.526. The latter are especially telling as these suggest that the model that most closely resembles the performance of subjects is one that is highly exploratory.

This should be interpreted with a grain of salt, however, because the "exploratory" algorithm's performance did not resemble that of subjects, as reported in the Results. In practice, subjects often chose a response alternative that was the "wrong" choice, according to the values of Q at that trial. Consequently, only a relatively low value of β, which would permit these "errors," could be an acceptable parameter.

As a result of these erroneous responses, the exploratory, best-fitting parameters were able to exceed chance performance in simulations of the task, both during training (by a hair) and testing (by a small amount). However, this above-chance performance was still far below what subjects were capable of, as plotted in Fig. 2 (A and E). The algorithm was also unable to fully capitalize on the beneficial reward information in the concordant gradient condition, making many errors during both training and testing and performing at levels well below that observed in subjects.

The inclusion of both the α and β parameters gives *Q*-learning flexibility, so although the exploratory parameters were the best-fitting, they were not necessarily the best that the algorithm could do in terms of total earnings. Consequently, we also implemented an exploitative algorithm ($\alpha = 0.15$, $\beta = 3.0$). This algorithm was very slightly above chance during reverse gradient training sessions and very slightly below chance during reverse gradient testing sessions, leading to a slightly lower return in those cases than the exploratory algorithm. However, during the concordant gradient phases, the exploitative algorithm performed near perfectly, and these added rewards more than made up for chance performance during the reverse gradient sessions.

The important takeaway of these simulations is twofold. On the one hand, neither the exploratory nor the exploitative algorithm is able to explain performance under the reverse gradient condition. Although the best-fitting parameters allow *Q*-learning to exceed chance, they do so by allowing the algorithm to make frequent exploratory choices against its better judgment. On the other hand, the very thing that makes the exploratory parameters effective in the reverse gradient case (frequent "errors") then puts a ceiling on how well it can perform in the concordant gradient condition. Meanwhile, the exploitation algorithm performs near perfectly in the concordant gradient condition, but this is expected because that condition can be solved in multiple ways. Thus, even if *Q*-learning were to dynamically adjust its β values from one session to the next, it still would not have enough flexibility to perform as well as subjects.

Statistical analysis: Binomial and logistic regression estimates of response accuracy

In Fig. 3, response accuracy was modeled on a session-by-session basis using binomial regression, with a logit link. Formally

$$\text{Count of Successes} = \textit{Binomial}(p, \text{Count of Attempts})$$

$$p = \frac{1}{1 + \exp(-\mu)} \quad (5)$$

$$\mu = \beta_{\varnothing} + \beta_{D} \cdot D + \beta_{R} \cdot R$$

In the model fit for each session, an intercept term $\beta_{\varnothing}$ was included, representing baseline response accuracy. In addition, performance was predicted in terms of both symbolic distance D and the reward magnitude of the correct alternative R, yielding two additional slope terms β_D and β_R. To minimize parameter covariance and accelerate numerical estimation, both D and R were centered by subtracting 2.5 from their empirical values. Thus, if subjects received only one drop of water for a

correct response to the pair AB in the reverse gradient condition, this would be coded as a reward of (−1.5) for the purposes of the regression. Centering in this way ensured that β_D and β_R reported the differential effects of distance and reward magnitude, and were expected to equal 0.0 when a variable had no influence. Note that because these were binomial regressions, they report the overall response accuracies for each session, without considering the time series of response accuracies within a session. In addition, the β_D term was omitted from the model entirely during the training phases, since all pairs during training had the same symbolic distance. Parameters were fit using multilevel models, yielding both population- and subject-level estimates, which were implemented in the Stan programming language (*15*).

Figure 4 also fit parameters using Eq. 5, but does so by pooling across the six phases of each experimental condition. For example, "training session 1 for the reverse gradient" in Fig. 4 incorporates the first sessions of phases 1, 3, 5, 7, 21, and 23. As in Fig. 3, estimates were computed session-wise as multilevel models. Because these estimates were based on six times as much data, their uncertainties are correspondingly smaller. The performance implications of the parameters described in Fig. 4 are then realized in Fig. 5, which shows a good correspondence between the observed empirical mean response accuracies for each pair (circles) and the model's estimates of performance (box-and-whisker plots).

The estimates from the binomial regressions in Figs. 3 to 5 give a good description of session performance overall. They also constitute good evidence that behavior was not based on expected reward, both because β_R overlapped with zero and because response accuracies were consistently higher than was predicted by *Q*-learning. However, these results do not, on their own, constitute evidence that a TI has occurred. This is because Eq. 5 fits performance for sessions overall, not at the first trial of each testing phase. In addition, Eq. 5 was fit using all the data, including the terminal items, which are often reported to have elevated response accuracy (*11*). Although Eq. 5 suggests that performance exceeded chance for all six critical pairs, that could reflect learning that took place during the first session of testing. The critical pairs (BD, CE, DF, BE, CF, and BF) were not part of the training set and did not include terminal items. One may conclude that TI occurred only if subjects exceeded chance on the six terminal items at the start of testing.

Figure 6 plots the results of three such critical tests. The first (Fig. 6, A and D) fits each of the six pairs using binomial regression. These estimates were performed simultaneously as a multilevel model and were based only on the first presentation of that stimulus pair; as such, these are estimates based on six observations per pair per subject. Unsurprisingly, the uncertainties in these estimates are large: The estimates are close to chance, but the whiskers envision a 95% credible range of values as low as 0.12 and as high as 0.88 in the reverse gradient condition.

To better leverage the available data, logistic regression was performed. This enabled estimation of response accuracy at "trial zero" for each pair, given how response accuracy to that pair evolved over the course of the first session. Formally, the full logistic regression model takes the following form

$$p(correct) = \frac{1}{1 + \exp(-\mu)} \tag{6}$$

$$\mu = \beta_\varnothing + \beta_t \cdot t + \beta_D \cdot D + \beta_R \cdot R$$

This adds one additional parameter, β_t, which predicts how response accuracy changes as a function of trial number t (where the first trial of testing is $t = 0.0$). As in previous analyses, the logistic regressions were performed as multilevel models, yielding both population-level and subject-level estimates.

In Fig. 6 (B and E), each of the six critical response pairs is fit in isolation from one another and separately for each condition, using only the $\beta_\varnothing$ and β_t terms from Eq. 6. Contrastingly, in Fig. 6 (C and F), all six critical trials are included, and both experimental conditions are pooled. As such, all of the parameters in Eq. 6 can be included. These parameter estimates are still uncertain (several pair estimates have credible intervals that overlap with zero), but they paint a compelling picture that, in general, the critical pairs exceeded chance in both conditions. To make a more precise statement about performance at transfer from training to testing, however, additional data are needed to refine these estimates.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/7/eaaw2089/DC1
Data and Analysis Scripts

REFERENCES AND NOTES

1. J. M. Keynes, *A Treatise on Probability* (Macmillan & Co., 1921).
2. B. W. Balleine, Sensation, incentive learning, and the motivational control of goal-directed action, in *Neurobiology of Sensation and Reward*, J. A. Gottfried, Ed. (CRC Press, 2011), chap. 13.
3. A. Rangel, C. Camerer, P. R. Montague, A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* **9**, 545–556 (2008).
4. A. San-Galli, S. Bouret, Assessing value representation in animals. *J. Physiol. Paris* **109**, 64–69 (2015).
5. D. L. Oden, R. K. R. Thompson, D. Premack, Can an ape reason analogically? in *The Analogical Mind*, D. Gentner, K. J. Holyoak, B. K. Kokinov, Eds. (Bradford, 2001), pp. 471–491.
6. D. C. Penn, K. J. Holyoak, D. J. Povinelli, Darwin's mistake: Explaining the discontinuity between human and nonhuman animal minds. *Behav. Brain Sci.* **31**, 109–130 (2008).
7. M. A. McDannald, F. Lucantonio, K. A. Burke, Y. Niv, G. Schoenbaum, Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* **31**, 2700–2705 (2011).
8. J. L. Jones, G. R. Esber, M. A. McDannald, A. J. Gruber, A. Hernandez, A. Mirenzi, G. Schoenbaum, Orbitofrontal cortex supports behavior and learning using inferred but not Cached values. *Science* **338**, 953–956 (2012).
9. R. P. Kesner, R. O. Hopkins, Mnemonic functions of the hippocampus: A comparison between animals and humans. *Biol. Psychol.* **73**, 3–18 (2006).
10. H. Terrace, The comparative psychology of ordinal knowledge, in *Oxford Handbook of Comparative Cognition*, T. R. Zentall, E. A. Wasserman, Eds. (Oxford Univ. Press, 2012), pp. 615–651.
11. G. Jensen, Serial learning, in *APA Handbook of Comparative Psychology: Perception, Learning, and Cognition*, J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, T. R. Zentall; American Psychological Association, Eds. (APA, 2017), pp. 385–409.
12. G. Jensen, F. Muñoz, Y. Alkan, V. P. Ferrera, H. S. Terrace, Implicit value updating explains transitive inference performance: The betasort model. *PLOS Comp. Biol.* **11**, e1004523 (2015).
13. O. F. Lazareva, E. A. Wasserman, Transitive inference in pigeons: Measuring the associative values of stimuli B and D. *Behav. Processes* **89**, 244–255 (2012).
14. M. Vasconcelos, Transitive inference in non-human animals: An empirical and theoretical analysis. *Behav. Processes* **78**, 313–334 (2008).
15. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
16. C. J. C. H. Watkins, P. Dayan, Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
17. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998).
18. R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis* (Wiley, 1959).
19. G. E. Schwartz, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
20. M. R. D'Amato, M. Colombo, The symbolic distance effect in monkeys (*Cebus apella*). *Anim. Learn. Behav.* **18**, 133–140 (1990).

21. G. Jensen, Y. Alkan, F. Muñoz, V. P. Ferrera, H. S. Terrace, Transitive inference in humans (*Homo sapiens*) and rhesus macaques (*Macaca mulatta*) after massed training of the last two list items. *J. Comp. Psychol.* **131**, 231–245 (2017).
22. F. R. Treichler, M. A. Raghanti, Serial list combination in monkeys (*Macaca mulatta*): Test cues and linking. *Anim. Cogn.* **13**, 121–131 (2010).
23. G. Jensen, D. Altschul, E. Danly, H. S. Terrace, Transfer of a serial representation between two distinct tasks by rhesus macaques. *PLOS ONE* **8**, e70285 (2013).
24. R. P. Gazes, O. F. Lazareva, C. N. Bergene, R. R. Hampton, Effects of spatial training on transitive inference performance in humans and rhesus monkeys. *J. Exp. Psychol. Anim. Learn. Cogn.* **40**, 477–489 (2014).
25. B. O. McGonigle, N. Chalmers, Are monkeys logical? *Nature* **267**, 694–696 (1977).
26. J. F. Cantlon, M. L. Platt, E. M. Brannon, Beyond the number domain. *Trends Cogn. Sci.* **13**, 83–91 (2009).
27. M. Ito, K. Doya, Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Curr. Opin. Neurobiol.* **21**, 368–373 (2011).
28. N. Tanner, G. Jensen, V. P. Ferrera, H. S. Terrace, Inferential learning of serial order of perceptual categories by rhesus monkeys (*Macaca mulatta*). *J. Neurosci.* **37**, 6268–6276 (2017).

Acknowledgments: We thank D. Freshwater, A. Gross, K. Liu, G. Spencer, and N. Tanner for assistance with data collection. **Funding:** This work was supported by U.S. National Institute of Mental Health grant numbers NIH-MH081153 and NIH-MH111703 awarded to V.P.F. and H.S.T. and by the Kavli Institute for Brain Sciences at Columbia University. **Author contributions:** G.J., V.P.F., and H.S.T. conceptualized the study, devised methodology, supervised volunteers, and wrote the original draft. G.J. and Y.A. collected data. G.J. analyzed data, developed task software, and created data visualizations. G.J., Y.A., V.P.F., and H.S.T. revised the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 28 November 2018
Accepted 21 June 2019
Published 31 July 2019
10.1126/sciadv.aaw2089

Citation: G. Jensen, Y. Alkan, V. P. Ferrera, H. S. Terrace, Reward associations do not explain transitive inference performance in monkeys. *Sci. Adv.* **5**, eaaw2089 (2019).