PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Superstitious learning of abstract order from random reinforcement

Yuhao Jin[a,b], Greg Jensen[b,c,d], Jacqueline Gottlieb[b,d,e,1,2], and Vincent Ferrera[b,d,e,1,2]

Edited by Charles Gallistel, Rutgers, The State University of New Jersey, Piscataway, NJ; received February 16, 2022; accepted July 1, 2022

Humans and other animals often infer spurious associations among unrelated events. However, such superstitious learning is usually accounted for by conditioned associations, raising the question of whether an animal could develop more complex cognitive structures independent of reinforcement. Here, we tasked monkeys with discovering the serial order of two pictorial sets: a "learnable" set in which the stimuli were implicitly ordered and monkeys were rewarded for choosing the higher-rank stimulus and an "unlearnable" set in which stimuli were unordered and feedback was random regardless of the choice. We replicated prior results that monkeys reliably learned the implicit order of the learnable set. Surprisingly, the monkeys behaved as though some ordering also existed in the unlearnable set, showing consistent choice preference that transferred to novel untrained pairs in this set, even under a preference-discouraging reward schedule that gave rewards more frequently to the stimulus that was selected less often. In simulations, a model-free reinforcement learning algorithm (*Q*-learning) displayed a degree of consistent ordering among the unlearnable set but, unlike the monkeys, failed to do so under the preference-discouraging reward schedule. Our results suggest that monkeys infer abstract structures from objectively random events using heuristics that extend beyond stimulus–outcome conditional learning to more cognitive model-based learning mechanisms.

learnability | randomness | superstitious learning | transitive inference | reinforcement learning

Significance

Past studies on learning and decision-making usually rely on the assumption that the task is learnable. However, humans and other animals often infer spurious relationships from coincidental associations, and it is unknown if this could be achieved without reward conditioning. Here, we exposed monkeys to sets of images that had a hidden hierarchical order and to unordered sets that lacked an underlying structure. Monkeys treated the unordered sets as if they had a hierarchical order even under reward schedules that incentivized random choices. The results cannot be explained by simple associative mechanisms that account for other types of spurious learning, suggesting that when presented with random events animals conjure elaborate model-based structures.

Learning is vital for survival. Learning mechanisms have been extensively studied in laboratory tasks that are learnable i.e., contain objective regularities that can be discovered through trial and error. However, in natural environments animals must not only learn but decide what to learn—that is, distinguish between true relationships that can be successfully learned and spurious associations that are random and unlearnable. For example, upon seeing a red Toyota driving by in the rain one may attempt to learn the mechanisms by which the Toyota prevents skids (a learnable question) but should not, ideally, attempt to learn if there is a relationship between red Toyotas and rain (a random unlearnable question) (1). However, humans (2–8) and nonhuman animals (9–12) learn spurious associations in a variety of conditions, making it unclear if they can reliably distinguish between learnable and unlearnable patterns.

Spurious learning is generally explained in terms of simple associative mechanisms that overestimate causal relationships between external events or between the animal's actions and outcomes (13–16). Indeed, simple associative learning models, like the Rescorla–Wagner model, depend heavily on correlations between events and can be easily fooled into strengthening associations based on coincidences (17). However, since animals can construct elaborate cognitive structures, it is an open question if they also inappropriately impose complex structures on objectively random events.

Here, we examined this question in the context of a "transitive inference" (TI) task that tested monkeys' ability to infer the ordinal relationships among a set of pictorial stimuli that had a hidden order. The task is well-suited to our question because it has been extensively characterized in multiple species (18–21) and shown to require mechanisms beyond reward associations (22–24). In our current task, monkeys viewed pairs of pictures drawn from an ordered, learnable set as in the classical TI task. In randomly interleaved trials, they viewed pairs from an unlearnable image set that had no hidden order and their choices were rewarded randomly. We analyzed whether choices on each stimulus set were consistent with a hidden order, how they were affected by different reward schedules, and if they could be reproduced by associative *Q*-learning algorithms.

Results

Subjects Learned Latent Order by Trial and Error. In each trial, rhesus monkeys ($n = 3$) saw two pictures and touched one to proceed. The pictures available for each

Author affiliations: [a]Department of Biological Sciences, Columbia University, New York, NY 10027; [b]Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027; [c]Department of Psychology, Reed College, Portland, OR 97202; [d]Department of Neuroscience, Columbia University, New York, NY 10027; and [e]Kavli Institute for Brain Science, Columbia University, New York, NY 10027

Author contributions: Y.J., G.J., J.G., and V.F. designed research; Y.J. performed research; Y.J. analyzed data; and Y.J., G.J., J.G., and V.F. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.G. and V.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jg2141@columbia.edu or vpf3@cumc.columbia.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2202789119/-/DCSupplemental.

Published August 23, 2022.

trial were both drawn from one of two sets of five pictorial stimuli. The stimuli in one set had an underlying rank order and subjects were rewarded for choosing the stimulus that had a higher rank in each pair (Fig. 1*A*, *Top*; the letter denotes the rank with A the highest and E the lowest rank). The stimuli in the other set were unordered and subjects were rewarded probabilistically regardless of their choices (Fig. 1*A*, *Bottom*; the letter only represents the serial number of each stimulus, but not rank). Because the objectively "correct" order of the stimuli could be learned from trial-and-error feedback for the former but not the latter set, we refer to the sets as, respectively, "learnable" (L) and "unlearnable" (U).

During each session, the subjects interacted with a new set of L and U stimuli. L and U trials were randomly interleaved and there was no explicit cue signaling which set the stimulus pair was drawn from. The sessions consisted of a training phase, during which subjects only experienced the four adjacent stimulus pairs from the L set and four randomly selected pairs from the U set, followed by a testing phase with all possible pairs (Fig. 1*B*). This design allowed us to examine if the subjects inferred an order during training and spontaneously transferred it to the new pairs during testing.

To better understand how subjects approached the U trials, each session used one of two different reward schedules for the U trials (while using the same schedule for L trials). Under the "preference-neutral" schedule (PN), the reward probability for U pairs was equated to that for L pairs by dynamically adjusting it to match the mean reward rate for L pairs on the preceding 10 L trials (while remaining independent of which U stimulus was chosen). Under the second, "preference-discouraging" reward schedule (PD), the reward probability for each U stimulus was inversely related to how recently the subject had selected it. The PD schedule thus discouraged repeated choice of any specific U stimulus and yielded maximal rewards if differences in U stimuli preferences were minimized and each U stimulus was selected equally (see *Materials and Methods* for details). Each subject finished 20 sessions for each schedule, with each session consisting of 25 to 50 training blocks (presenting the training pairs for each set) and 10 testing blocks (presenting all pairs). Consistent with the reward schedule design, the reward rates for the U and L sets were indistinguishable for the PN schedule (L: 0.7581 ± 0.0107, U: 0.7473 ± 0.0109) but differed appreciably for the PD schedule (L: 0.7801 ± 0.0121, U: 0.4626 ± 0.0035) (*SI Appendix*, Fig. S1*A*).

As expected from previous studies, all the subjects reliably learned the order of the L sets, shown by above-chance response accuracies and by a robust symbolic distance effect (SDE), whereby reward rates increased as a function of the difference in rank between the two stimuli presented on each trial. The SDE for L sets was significant on average (Fig. 1*C*, *Left*; PN: $F(3,236) = 93.391$, $p < .001$; PD: $F(3,236) = 76.34$, $p < .001$) and in each individual monkey (*SI Appendix*, Fig. S1*B*). In contrast, there was no significant SDE for the U set, as expected given that SD was calculated as the difference in arbitrary serial numbers assigned to the stimuli (Fig. 1*C*, *Right*; PN: $F(3,236) = 0.747$, $p = .524$; PD: $F(3,236) = 0.917$, $p = .432$; *SI Appendix*, Fig. S1*B*).

Subjects Chose As If Imposing an Order on Unordered Stimuli. Despite the lack of objective order among the U stimuli, subjects seemed to treat them as if they were ordered. Fig. 2*A*, *Top* shows this result for an example session with the PN reward schedule, in which the subject developed a clear choice preference in the L set (which was consistent with the objective stimulus

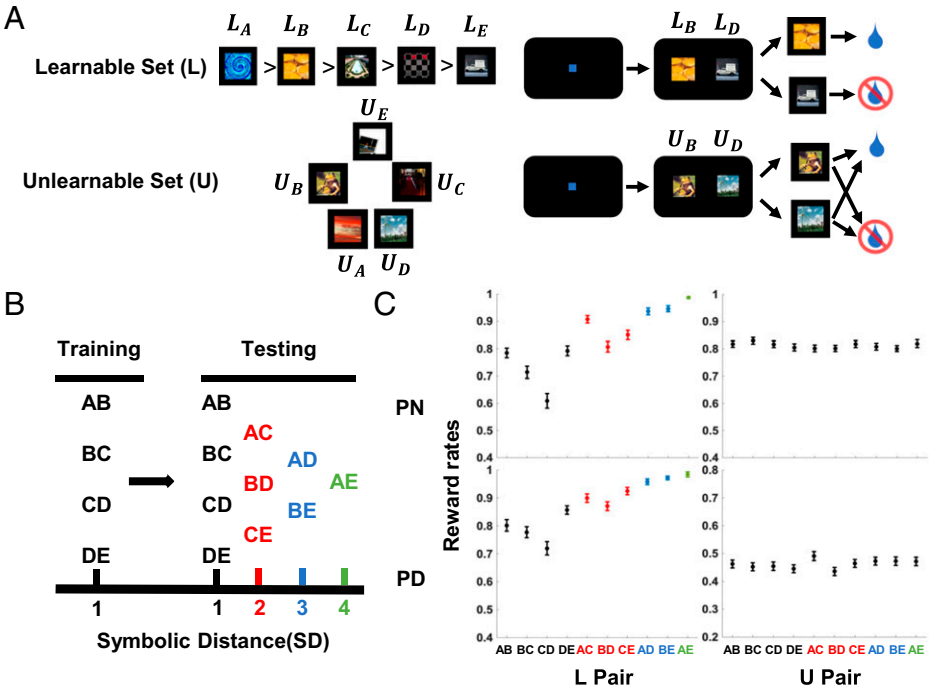

Fig. 1. Task paradigm. (*A*) Subjects were tasked with discovering the implicit ordering of sets comprising five pictorial stimuli. In "learnable" sets (L), the stimuli were assigned an order that could be inferred by trial and error. In learnable trials, the picture–outcome association was consistent and predictable: The picture with a higher rank was always associated with reward, whereas the other was not (*Top*). In the "unlearnable" sets (U), there was no predefined order and feedback was delivered probabilistically. In unlearnable trial, either response could potentially result in reward. Under the preference-neutral (PN) condition, reward probabilities for U pairs were yoked to recent performance on L pairs; under the preference-discouraging (PD) condition, the reward probability for each stimulus was inversely related with how recently it had been selected. See *Materials and Methods* for details. (*B*) Subjects were first presented with training blocks consisting of only the adjacent pairs (SD = 1). After training, they transitioned to testing blocks consisting of all the possible pairs (SD = 1 to 4). The transition from training to testing let us evaluate performance on novel pairs that rely on TIs. (*C*) Reward rates over all the L and U pairs across subjects. Error bars denote the SEMs.

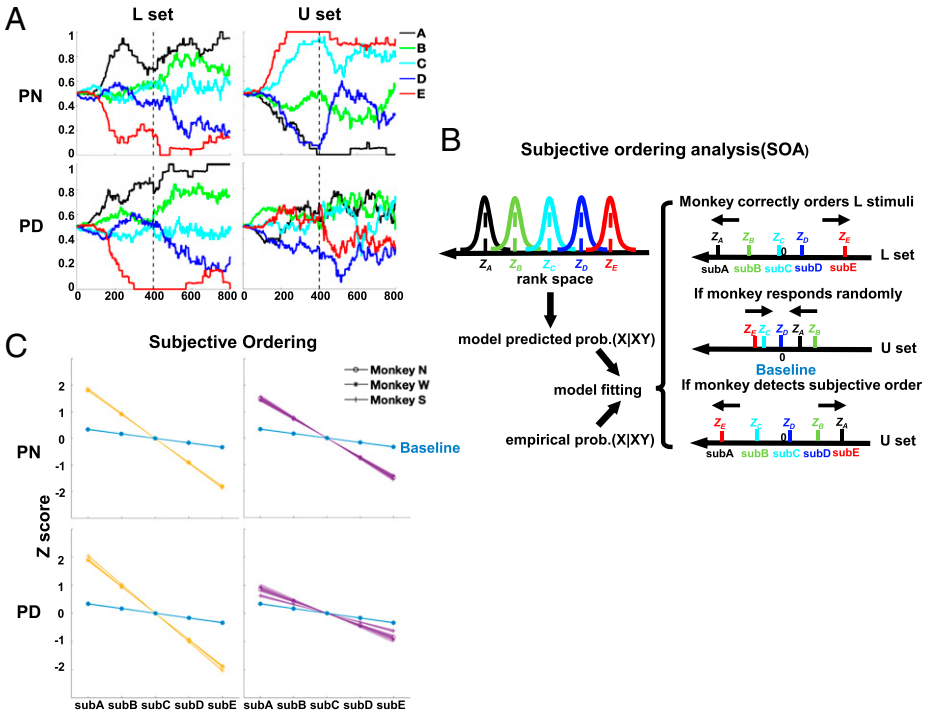

Fig. 2. Subject created abstract ordering upon the unlearnable stimuli. (*A*) Example session showing moving averages across trials ($n = 80$ for training and $n = 50$ for testing) of choice frequency for over L and U stimuli. (*B*) A schematic depiction of the SOA. The model assumed that a subject represented each stimulus as a position along a linear continuum with normally distributed uncertainty. Each stimulus had its own mean (*Z*-score) and SD parameter. The model was used to infer the most probable *Z*-scores for each stimulus given the subject's history of choices across all pairs. All stimuli could be subjectively labeled subA to subE based on their *Z*-scores. The L set was expected to display a subjective order that was consistent with the true order (*Top*). If a subject selected U set items randomly, the inferred *Z*-scores would shrink together close to zero (*Middle*). However, if a subject created a subjective order, then the *Z*-scores would spread out to reduce their overlapping uncertainties (*Bottom*). Baseline estimates for the *Z*-scores were obtained by simulating random responding, in order to have a rigorous null against which to compare behavior. (*C*) Estimated subjective item positions for both L and U sets under both reward schedules during the testing phase. All subjects showed stronger preference orderings (i.e., steeper slopes) than baseline. Error bars represented the SEMs.

label) and also a clear preference within the U set (despite the arbitrary stimulus labels). Under the PD schedule, a similar result for the L set was found, and the U set still manifested differential preferences over stimuli (Fig. 2*A*, *Bottom*).

To estimate the strength of these apparent preferences among U stimuli and test if they could occur by chance, we used a model-based subjective ordering analysis (SOA). The analysis fit choices in each session based on the assumption that subjects represented stimuli along a linear continuum with some uncertainty about stimulus positions (Fig. 2*B*, *Materials and Methods*, and ref. 22). For data acquired during the testing phase of each session, the analysis produced a *z*-score indicating the relative rank of each stimulus. A stronger *z*-score gradient indicated stronger preference, more consistent choices, and less overlap between inferred stimulus ranks. For the L sets, the gradients (slopes) over the *z*-scores were significantly higher than would be expected from a baseline of random responding during the testing phase (PN: all $p < .001$; PD: all $p < .001$, rank-sum test; Fig. 2*C*, *Left*). Furthermore, the gradients for the L sets were equivalent whether stimuli were ranked according to the objective ordering defined by the experiments or according to the subjective ordering estimated by the analysis (testing phase: PN: all $p > .2$; PD: all $p > .7$, rank-sum test). Put another way, when an objective ordering existed, the SOA of behavior reliably recovered the true stimulus ranks from each subject's preferences, confirming the validity of the SOA.

The *z*-score gradients for U sets for all three subjects displayed slopes that were significantly steeper than baseline in both PN and PD schedules (Fig. 2*C*, *Right*; PN: all $p < .001$; PD: all $p < .001$, rank-sum test) and were stronger in the PN relative to the PD schedule (all $p < .001$, rank-sum test). This suggests that subjects displayed consistent preferences among U stimuli, despite receiving rewards that were independent of stimulus in the PN schedule, or actively discouraged preferences in the PD schedule. Evidence for subjective ordering among U stimuli was also found during the training phase of each session (*SI Appendix*, Fig. S2*A*). The strength of the subjective ordering did not systematically change across sessions (*SI Appendix*, Fig. S2*B*) and was not correlated with the gradient in L sets (*SI Appendix*, Fig. S2*C*), suggesting these preferences are stable over the long term and are not explained by general engagement with the task. Thus, the monkeys' preferences appeared to reveal a tendency to impose order on unordered stimulus sets, even under a reward schedule in which such preferences incurred a cost by reducing the rate of reward.

Subjects Transferred the Subjective Ordering from Training to Testing. Comparisons of training and testing stages for the U sets showed that the subjective preferences that developed during training remained consistent during the testing stage. Fig. 3*A* illustrates this result for a representative subject under PN and PD schedules, in which the rank order preference for stimuli during testing (shown by the color labels) was the same as the rank ordering during training (shown by the relative position of the traces). Two analyses verified this result quantitatively. First, the *z*-scores over each U stimulus estimated from training and testing data were significantly correlated for the PN schedule (*SI Appendix*, Fig. S3*A*, $r > 0.59$, $p < .001$) and for two of three subjects in the PD schedule (*SI Appendix*, Fig. S3*A*, $r > 0.18$, $p < .001$). Second, during testing on the U sets, subjects showed a robust effect of

subjective symbolic distance (subSDE)—i.e., symbolic distances coded based on the subjective ordering inferred during training)—that was similar to the objective symbolic distance effect for the L set (where distances were coded based on objective order; *SI Appendix*, Fig. S3*B* and Fig. 3*B*). This further supports the idea that subjects were choosing as if there were a consistent order and thereby showed stronger subjective bias on pairs with larger subSD. A logistic regression during the testing stage revealed slopes for subSDE that were significantly higher relative to a randomized control for each subject in the PN schedule (Fig. 3*B*, *Top Right*; U 95% CI for the subSD regressor N: [0.3675, 0.3707], W: [0.4065, 0.4100], S: [0.4564, 0.4604]; subSD regressor shuffled N: [0.0010, 0.0040], W: [0.0421, 0.0452], S: [−0.0216, −0.0182]) and for two of three subjects in the PD schedule (Fig. 3*B*, *Bottom Right*; U 95% CI for the subSD regressor N: [0.1588, 0.1618], W: [0.1298, 0.1328], S: [0.0154, 0.0183]; shuffled N: [−0.0437, −0.0408], W: [0.0298, 0.0327], S: [−0.0372, −0.0343]). Thus, subjective orderings estimated from training could be used to make predictions about performance for novel U pairs presented during the testing phase.

Q-Learning Does Not Explain Observed Behavior. A difficulty in evaluating behavior under an ostensibly "uninformative" condition is that subjects may draw spurious conclusions from the random feedback that was provided. Put another way, reinforcement learning (RL) models that depend on experiential reward can sometimes form preferences even when feedback is random and uninformative. To evaluate this possibility, we simulated behavior using a model-free *Q*-learning RL algorithm (25), which is often considered a canonical example of reward prediction error learning. In addition to having a "learning rate" parameter, our *Q*-learning implementation used a softmax decision rule (26, 27) with a "temperature" parameter that governed how much random variation was introduced into decisions. Posterior parameter distributions for each subject were estimated numerically using the Stan programming language (28, 29), as described in *Materials and Methods*. As expected, *Q*-learning succeeded in learning the veridical order in L set and showed matching reward rates between L and U sets under the PN schedule but decoupled reward rates under the PD schedule (*SI Appendix*, Fig. S4*A*).

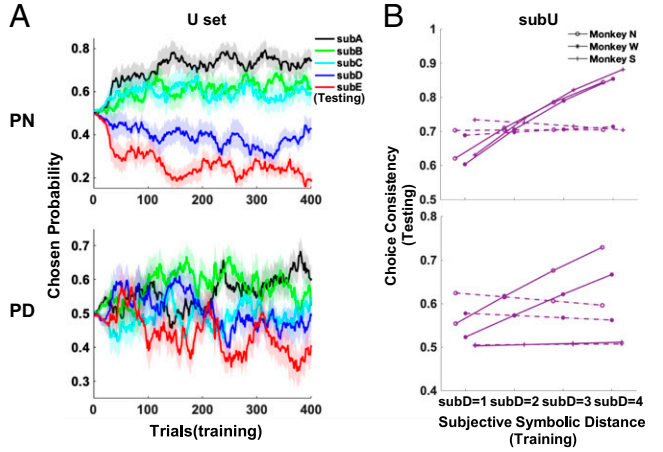

Fig. 3. Subject transferred the subjective order on the U set from training to testing. (*A*) Moving averages across trials (n = 16) of choice frequency of each U stimulus during training for an example subject. Here, the subA–subE labels were determined using the testing data. Shaded regions correspond to the bootstrapped 95% CI. (*B*) Logistic regression estimates of preference as a function of symbolic distance between the subjective ranks of U stimuli at the start of testing. Solid lines denote subject estimates, and dashed lines represent a null case with the distance shuffled. Error bars denote the SEs over all the posterior Bayesian regression coefficients. However, the error bars are invisible because they are smaller than the size of the lines.

While *Q*-learning developed preferences (and thus a subjective ordering) among U stimuli when rewarded using the PN schedule, it failed to display reliable preferences under the PD schedule. Under the PN schedule, applying the SOA to the model choices showed that *Q*-learning produced significant subjective ordering for the U sets relative to baseline of random response, in a manner similar to the monkeys (Fig. 4*A*, averaged behavior: top, mean slope baseline: −0.1671; *Q*-learning: −0.7314; 95% CI Monkey: [−0.7503, −0.7375], *SI Appendix*, Fig. S4*B*, *Top*). Applying the transfer analysis showed that *Q*-learning produced significant transfer from training to testing, albeit to a lesser extent than shown by the monkeys (Fig. 4*B*, averaged behavior: top, U subSD regressor 95% CI Monkey: [0.4095, 0.4143], mean *Q*-learning: 0.2234; subSD regressor shuffled 95% CI Monkey: [0.0068, 0.0107], mean *Q*-learning: 0.0092, *SI Appendix*, Fig. S4*C*, *Top*). However, under the PD schedule, *Q*-learning performed worse than baseline in forming subjective orderings (Fig. 4*A*, averaged behavior: bottom, mean slope baseline: −0.1671; *Q*-learning: −0.1211; 95% CI Monkey [−0.4058, −0.3916], *SI Appendix*, Fig. S4*B*, *Bottom*) and showed no transfer effect (Fig. 4*B*, averaged behavior: bottom, U subSD regressor 95% CI Monkey: [0.0997, 0.1059], mean *Q*-learning: −0.0026; subSD regressor shuffled 95% CI Monkey: [−0.0176, −0.0135], mean *Q*-learning: 0.0028, *SI Appendix*, Fig. S4*C*, *Bottom*). Therefore, our model simulations showed that the subjective ordering is unlikely to be due to experienced reward and instead may rely on mechanisms that extend beyond standard model-free *Q*-learning.

Discussion

While spurious learning has been intensively studied, it is usually explained by associative learning and little is known about whether animals superstitiously infer more complex relationships from random events. Here we provide evidence that, in an unstructured environment, animals do not respond randomly but act as if there was a structure. We exposed nonhuman primates to unordered sets of stimuli within the context of a TI paradigm and discovered that they treated the sets as though they were ordered. Subjects developed preferences consistent with the stimuli's being rank-ordered, and these "subjective orderings" persisted throughout the session and were predictive of choices made to new pairs.

Importantly, we show that subjective orderings were inferred through mechanisms that go beyond simple associative learning. For unlearnable sets, in stark contrast with learnable sets, subjects were insensitive to past feedback and showed equal probability of win–stay and lose–stay strategies (*SI Appendix*, Fig. S5 and refs. 5, 6, and 11). Moreover, choices consistent with subjective ordering remained strong in a PD schedule that actively disrupted them by dynamically increasing reward probability for whichever alternative had been selected least often. In contrast, a *Q*-learning algorithm failed to show subjective ordering under this schedule, although it replicated it under a PN schedule that did not discourage consistent preferences, capturing the well-known vulnerability of associative models to spurious reward correlations (13–17). These results are consistent with a wealth of studies showing that pure associative learning is not sufficient to explain TI learning (22–24, 30). Importantly, they show that monkeys ascribe subjective structure to objectively random events based on inferential mechanisms that rely on more complex assumptions than retrospective reward maximization.

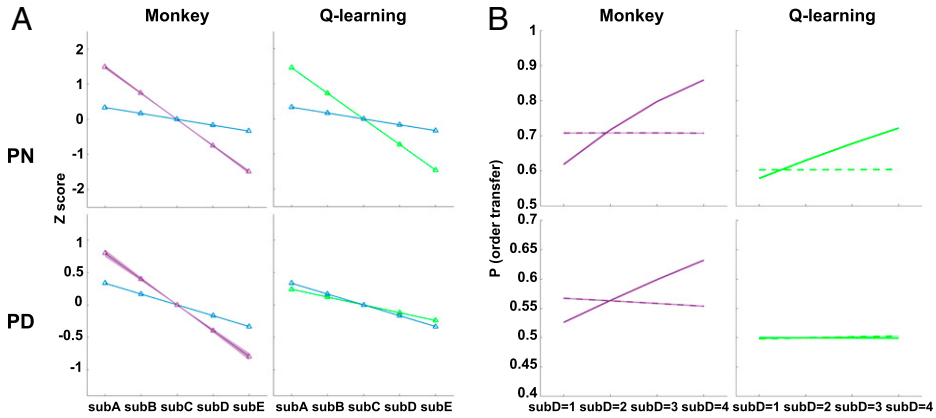

Fig. 4. *Q*-learning cannot explain monkeys' behavior (*A*) Subjective ordering analysis of RL simulations using *Q*-learning, as compared with that of subjects. (*B*) Mean performance at the start of testing, sorted by symbolic distance (solid lines) and a null case (dashed lines). Error bars denote the SEs over all the posterior Bayesian regression coefficients.

The mechanisms generating subjective ordering are unknown. One possibility is that subjective preferences rely on a memory buffer that retains a record of recent items and reward history, consistent with the proposal that a working memory buffer encodes sequences and serial order (31, 32). However, these approaches track reward value similar to associative learning, making it unlikely that they would perform better than the *Q*-learning algorithms we tested (as both mechanisms track reward value, with the slight difference being whether such value represents the transient or accumulated reward history).

A second possible mechanism involves generalization. Encouraged by the fact that U and L stimuli were unsignaled and randomly interleaved in our task, subjects may have extrapolated from their experience with serial learning and learnable trials and treated all the sets they experienced as having a learnable order. If subjects begin the task believing that each stimulus is preassigned to a certain rank, it stands to reason that this a priori representation would be resilient against at least some counterfactual information. This view is consistent with other studies in which monkeys and humans were shown "derived" pairings that mixed stimuli from two different pretrained objectively ordered sets. In such tasks, subjects rely on the known ranks held by each stimulus in their original sets to judge the novel across-set combinations, suggesting that they spontaneously assume that the two sets use the same ranking scale even when there is no logical necessity for this being the case (33–35).

More broadly, this view is consistent with the proposed role of generalization heuristics in guiding exploration in complex contexts under high uncertainty (36–38). Popular methods for coding generalization in RL algorithms involve inferring an abstract latent state φ from one context and applying it to guide learning in a new context (39). In our case, since each state (each trial) is independent of every other, generalization may be operationalized with the equation $E(R|s,a) = E(R|\varphi(s),a)$, where $E(R|s, a)$ is the expected reward given state s and action a and φ refers to an internal ranking over all stimuli analogous to those proposed in positional inference models in the previous TI literature (35).

A third, not mutually exclusive, mechanism is related to the idea that conflict, uncertainty, and ambiguity are aversive and pose cognitive costs (40–42). In humans, choices among food items with similarly high value are associated with higher anxiety relative to choices among more distinct items (43), providing direct evidence that making decisions under higher uncertainty has affective costs. Should animals attempt to avoid these costs by reinterpreting the situation, this could result in irrational behaviors like illusions of causality and superstitious choices (44, 45). Thus, our monkeys' behavior may have been motivated by a desire to reduce the conflict they experienced between the subjective order they inferred and the reward feedback they received that was random and unrelated to that order. This process may be modeled using variations in learning rates like those proposed by the Pearce–Hall model (46) or more recent implementations (47), by postulating that, when animals receive feedback that is inconsistent with an inferred structure, this suppresses learning rates, leading to superstitious learning that is insensitive to reward outcomes. Because this mechanism assumes that animals derived a subjective order, it would complement rather than replace a generalization-based mechanism that potentially gave rise to this order.

We note, however, that while the above mechanisms may reproduce the fact that fictitious learning occurs, they do not necessarily account for the specific order that a subject adopts for a specific stimulus set. Predicting this feature of our data will require an account of how animals derive intrinsic preferences, as has been recently attempted (48, 49). Thus, the relative roles of generalization, attitudes to uncertainty, and intrinsic preferences in motivating inferences of spurious structures will be important areas for future investigations.

Whatever the eventual mechanism turns out to be, our result that subjective ordering persisted through the PD schedule, which reduced reward rates, suggests that it is a powerful phenomenon that may lead to suboptimal choices. This may be particularly important in "strategic learning" scenarios, in which learners must decide how to allocate time and effort among competing learning activities. A recent report showed that humans devoted disproportionate effort to random and unlearnable tasks at the expense of improving on alternative learnable tasks (8). A theoretically efficient method for avoiding "randomness traps" is to value competing activities in proportion to learning progress—the extent to which one's success rates improve over time (1, 50). However, this strategy may be considerably weakened if people, similar to the monkeys we studied, are less sensitive to actual reward rates—or, more precisely, to the contingency between their choices and outcomes—and act based on their assumption that a structure exists. Future work could further probe the mechanisms of superstitious learning and allow participants to freely choose between learnable and unlearnable trials to test whether they, as our results hint, exert disproportionate effort to learning superstitious structure from random feedback.

Materials and Methods

Subjects. Subjects were three adult male rhesus macaques (*Macaca mulatta*), N, W, and S. All subjects had different amounts of prior experience with the classic TI task (single ordered set, transfer paradigm) from years (N and S) to only a month (W). However, none of the subjects had been exposed to the unordered U sets, or to the PN and PD schedules. Subjects were water-restricted for maintaining high motivation. Subjects earned the reward by receiving water drops, each drop having a volume of about 0.1 mL. Typical performance per session yielded between 150 mL and 300 mL. Subjects were also provided a ration of biscuits each day before the task and fruit as an extra bonus after the task. The study was implemented obeying to the guidelines provided by the *Guide for the Care and Use of Laboratory Animals* of the NIH. This work was also approved by the Institutional Animal Care and Use Committees at Columbia University.

Apparatus. Subjects performed the task by using touchscreen connected to a computer (Windows 10) while sitting on the chair. The touchscreen (Elo Touch Solutions) presented subjects with a 15- × 12-inch high-definition display (1,280 × 1,024 resolution at 60 Hz); both showed the objects and recorded the response. Tasks were programmed in MATLAB (version 2018a; MathWorks) using Psychophysics Toolbox (51). To deliver fluid rewards, the computer sent out the command to the Arduino Uno interface, which then relayed the signal to the solenoid valve with 0.1 mL of water or juice being delivered through a tube installed on the primate chair.

Procedure. Pictorial stimuli were selected at random from a large bank of stock photographs and further processed to equalize their size (250 × 250 pixels). Sets were examined in advanced to confirm that they did not include stimuli that could be easily confused for one another. Each day, subjects performed one session and presented with two pictorial sets with five stimuli each. Different sets were used over days and across subjects. Over the two sets, one set is learnable (L), meaning the stimuli were preassigned an arbitrary rank order. To earn rewards, the subject was required to learn to infer the veridical order by trial and error (denoted as L_A L_B L_C L_D L_E). Another set is unlearnable (U), meaning the stimuli were not ordered, and the subject would acquire zero knowledge of order because of the random feedback (denoted as U_A U_B U_C U_D U_E; Fig. 1*A*).

During each trial, a pair of stimuli was shown side-by-side on the touchscreen. Both stimuli were drawn either from the L set or the U set. The subject touched one of the stimuli to indicate their choice. At the beginning of each trial, a solid blue square (100 × 100 pixels) was presented at the center of the screen to attract the subject's attention. The subject was allowed 3 s to touch the square to initiate the trial; otherwise, the current trial would be aborted, and the same trial would be presented again until a response was made. After the initiation, the blue square disappeared and the randomly drawn stimulus pair was shown with each picture at an equal distance (289 pixels) from the center. The subject had 4 s to make a decision by touching one of the stimuli; otherwise, the trial would be skipped and missed forever and the task would move on to the next trial. If the pair came from the L set, the subject would receive positive feedback (green check sign, followed by two drops of reward and sound cue) or negative feedback (red X sign, followed by 3-s time-out with the screen being dark) based on whether the decision accorded with the order in the L set. Therefore, choosing the stimulus with higher rank always resulted in a positive outcome. If the current stimulus pair came from the U set, the feedback would be delivered probabilistically independent of which pair of stimuli was displayed.

Each subject was exposed to two schedules that kept the design for L set the same but varied for the U set with the same number of sessions (20 per subject for each schedule). Under the PN schedule, the reward was independent of which stimulus was chosen. Reward probability for each U trial was yoked to the average reward rates over the 10 past L trials to the current trial.

Under the PD schedule, the reward probability over all stimuli started from 0.3 when each was presented the first time but was inversely correlated with the recent chosen frequency over all the subsequent presentations by following the equation $\left(1-(1-0.3)^{n+1}\right)$, where n denotes the number of trials that the stimulus remained unchosen since the last time it was chosen. Therefore, the degree of preference over all U stimuli did not affect the reward rates under the PN schedule since it is totally dependent on the subject's performance on the L trials, whereas the PD schedule punished persistent preference for any U stimulus because any repetitive selection of the same stimulus would result in lower reward rates than choosing each stimulus equally.

Each session comprised multiple adjacent pair training blocks (N and S: 25; W: 50), followed by multiple all-pairs testing blocks (all subjects: 10). Each training block contained 16 trials (4 different stimulus pairs per set [AB, BC, CD, DE] × 2 stimulus sets × 2 for spatial counterbalancing) and each testing block contained 40 trials (10 different stimulus pairs from each set × 2 stimulus sets × 2 for spatial counterbalancing) (Fig. 1*B*). Within each block, the sequence of the trials was randomized with L and U trials interleaved to ensure that subjects could not predict the future trials. Overall, N and S completed up to 800 trials per session, whereas W completed up to 1,200 trials.

Data Analysis.

SOA. To best infer the subjective ordering over all L/U stimuli and evaluate the ordering strength, we assumed a model-based representation where each stimulus i normally distributed along the same positional continuum with its own mean μ_i and variance σ_i. When a given pair (XY) was presented, the stimulus with higher mean was more likely to be chosen. For example, the probability to choose X was given by

$$p(X|XY) = \frac{\theta}{2} + (1-\theta)\int_0^\infty \mathcal{N}(x|\mu_X - \mu_Y, \sqrt{\sigma_X{}^2 + \sigma_Y{}^2})dx. \quad [1]$$

Here, θ denotes the degree to which subject ignored the current presentation and made random responses. Thus, subjects had probabilities that could be computed simultaneously for four pairs (during training) or 10 pairs (during testing) per session in each schedule [we only considered $p(AB)$ but not $p(AB)$ since $p(AB) = 1 - p(AB)$, and so on, for other pairs]. These simultaneous equations were solved using Bayesian multilevel model fitting using Markov chain Monte Carlo (MCMC) method in the Stan programming language (28). For more details on this procedure see ref. 22.

After model fitting, each parameter μ_i was Z-scored relative to other position estimates; that is, the mean value of all stimuli in a set was centered at zero. The subjective ordering was manifested by performing linear regression over the sorted Z-scores in descending order, by which we could use the slope to measure the ordering strength and sorted Z-scores to label the subA–subE over each set of L/U stimuli. Since even random responses are expected to display a nonzero slope due to the sorting step, we simulated sessions in which each subject responded entirely randomly. Then, each session of the simulated data went through the aforementioned subjective ordering procedure, which gave us the "baseline" slopes that were expected from this null model of random responding, in order to better evaluate how far that the subjective ordering is from random responding.

Logistic regression. To look at whether subject carried the subjective ordering from training to testing stage, we applied the following logistic regression model:

$$p(\text{Consistent to the ordering}) = (1 + \exp(-\mu))^{-1}$$
$$\mu = \beta_0 + \beta_t \cdot t + \beta_D \cdot D + \beta_{tD} \cdot t \cdot D. \quad [2]$$

In the regression model, the response variable represented whether subject's or artificial agent's choice during testing accorded to the veridical order in L set or the subjective order in U set during training (Boolean output). Such dependent variable indicating ordering transfer was predicted by trial number (t), symbolic distance (D, based on ground truth order for the L set and subjective ordering for the U set), and their interaction, yielding three slope terms β_t, β_D, and β_{tD}. The intercept term β_0 and β_D served to estimate p(*Consistent to the ordering*) at trial zero with respect to different D for both sets. Therefore, typical symbolic distance effect would show a significant positive β_D. In other words, a larger value of β_D corresponds to a preference that is more consistent to the order from training, and thus that the subjective order transferred from training to testing. Both t and D were centered at zero to decrease the slope covariances before Bayesian multilevel model fitting performed in the Stan programming language using the MCMC method (28). Additionally, a separate analysis was performed in which D was shuffled over trials, sessions, and subjects for both L and U sets in order to provide a control case and help evaluate the significance of the SDE.

Q-learning simulation. We used a model-free RL algorithm *Q*-learning to examine the hypothesis that the subjective ordering could be solely driven by evaluation of each stimulus's reward value under the random feedback. Since each trial

was totally independent of each other, the agent's action on trial t (a_t) had no effect on the next state S_{t+1}. Therefore, the Q-learning model implied for both sets were identical to the Rescorla–Wagner model (52):

$$\text{Value Updating Rule :} \quad Q(a_t|s_t) \leftarrow Q(a_t|s_t) + \delta[c_t - Q(a_t|s_t)] \qquad [3]$$

$$\text{Action Policy :} \quad p(O_t|s_t) = \text{softmax}(Q, T) = \frac{exp(Q(O_t|s_t)/T + I_t * \gamma)}{exp(Q(O_t|s_t)/T) + I_t * \gamma) + exp(Q(N_t|s_t)/T)}, T \geq 0. \qquad [4]$$

In the model, s_t was the stimulus pair in each trial, a_t was the agent's choice, c_t was the actual reward that agent received, the probability that subject chose certain option O_t over the unchosen one N_t was calculated by the softmax function. The Q-learning model implemented here followed the asymmetrical updating rule assuming that only the Q value of the chosen stimulus got updated. Three hyperparameters, δ (learning rate, determines how fast the value gets updated), T (temperature, decides how much variability to introduce into choices), and γ (spatial bias, $I = +1$ for left response and -1 for right response), underwent Bayesian multilevel model fitting performed in the Stan programming language using the MCMC method (28). The mean values of the best-fitting parameters were shown as follows: PN: L set δ: N 0.0097 W 0.0036 S 0.0155; T: N 0.0788 W 0.0527 S 0.0671; γ: N 0.3713 W −0.0879 S 0.4620; U set δ: N 0.0152 W 0.0082 S 0.0280; T: N 0.0726 W 0.0639 S 0.0785; γ: N 0.5957 W −0.1008 S 1.1292; PD: L set δ: N 0.0125 W 0.0040 S 0.0108; T: N 0.0849 W 0.0445 S 0.0413; γ: N 0.1437 W −0.6379 S 0.7093; U set δ:N 0.4601 W 0.5285 S 0.3048; T: N 15.8807 W 22.1249 S 0.9469; γ: N 0.5722 W −0.6415 S 2.3307. The parameters indicated that for the U set temperature was much higher under the PD (T close to or higher than 1) than under the PN schedule (T smaller than 1), indicating that subjects' decisions relied less on the retrospective reward history because of the high volatility of the feedback under the PD schedule. Next, we simulated the Q-learning behaviors with 1,500 sessions per subject under each schedule, which were thereby used for SOA and logistic regression. It was noteworthy that the response variable no longer referred to single trial (binary outcome) but rather to a percentage of trials consistent to the order out of a collection of trials pooled over all sessions that shared the same regressor value (t, D, or both). Such binomial regression would largely lower the computational cost.

Past-trial effect analysis. To evaluate how a subject's current decision was adaptive to the past-trial feedback, we quantified the percentage of win–stay and lose–stay for each L and U pair separately from the testing stage session by session for each subject and schedule, calculated as the chance of staying on the stimulus in pair XY that was rewarded/punished when last time XY was presented. Finally, we averaged the percent of win–stay and lose–stay over all L and U pairs.

Data, Materials, and Software Availability. Monkeys' behavior data, model fitting, and simulation results and the analysis codes sufficient to generate all the figures have been deposited in Figshare (https://figshare.com/articles/dataset/Supplementary_data_package_zip/19175612) (53).

ACKNOWLEDGMENTS. We thank Dr. Allain-Thibeault Ferhat and Yvonne Li for their feedback on early drafts of the manuscript. This work was supported by grant NIH-R01MH111703 from the NIH (V.F.).

1. J. Gottlieb, P. Y. Oudeyer, M. Lopes, A. Baranes, Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends Cogn. Sci.* **17**, 585–593 (2013).
2. S. Ahmad, H. Huang, A. J. Yu, Cost-sensitive Bayesian control policy in human active sensing. *Front. Hum. Neurosci.* **8**, 955 (2014).
3. J. S. Ide, P. Shenoy, A. J. Yu, C. S. Li, Bayesian prediction and evaluation in the anterior cingulate cortex. *J. Neurosci.* **33**, 2039–2047 (2013).
4. A. J. Yu, J. D. Cohen, Sequential effects: Superstition or rational behavior? *Adv. Neural Inf. Process. Syst.* **21**, 1873–1880 (2008).
5. A. Abrahamyan, L. L. Silva, S. C. Dakin, M. Carandini, J. L. Gardner, Adaptable history biases in human perceptual decisions. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3548–E3557 (2016).
6. E. Bosch, M. Fritsche, B. V. Ehinger, F. P. de Lange, Opposite effects of choice history and evidence history resolve a paradox of sequential choice bias. *J. Vis.* **20**, 9 (2020).
7. H. Matute *et al.*, Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Front. Psychol.* **6**, 888 (2015).
8. A. Ten, P. Kaushik, P. Y. Oudeyer, J. Gottlieb, Humans monitor learning progress in curiosity-driven exploration. *Nat. Commun.* **12**, 5972 (2021).
9. B. F. Skinner, 'Superstition' in the pigeon. 1948. *J. Exp. Psychol. Gen.* **121**, 273–274 (1992).
10. T. R. Zentall, When humans and other animals behave irrationally. *Comp. Cogn. Behav. Rev.* **11**, 25–48 (2016).
11. D. Lee, M. L. Conroy, B. P. McGreevy, D. J. Barraclough, Reinforcement learning and decision making in monkeys during a competitive game. *Brain Res. Cogn. Brain Res.* **22**, 45–58 (2004).
12. T. Teichert, V. P. Ferrera, Suboptimal integration of reward magnitude and prior reward likelihood in categorical decisions by monkeys. *Front. Neurosci.* **4**, 186 (2010).
13. H. Matute, I. Yarritu, M. A. Vadillo, Illusions of causality at the heart of pseudoscience. *Br. J. Psychol.* **102**, 392–405 (2011).
14. C. Orgaz, A. Estévez, H. Matute, Pathological gamblers are more vulnerable to the illusion of control in a standard associative learning task. *Front. Psychol.* **4**, 306 (2013).
15. E. Daprati, A. Sirigu, M. Desmurget, D. Nico, Superstitious beliefs and the associative mind. *Conscious. Cogn.* **75**, 102822 (2019).
16. K. R. Abbott, T. N. Sherratt, The evolution of superstition through optimal use of incomplete information. *Anim. Behav.* **82**, 85–92 (2011).
17. F. Blanco, H. Matute, M. A. Vadillo, Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learn. Behav.* **41**, 333–340 (2013).
18. B. O. McGonigle, M. Chalmers, Are monkeys logical? *Nature* **267**, 694–696 (1977).
19. H. Davis, Transitive inference in rats (*Rattus norvegicus*). *J. Comp. Psychol.* **106**, 342–349 (1992).
20. A. B. Bond, A. Kamil, R. P. Balda, Social complexity and transitive inference in corvids. *Anim. Behav.* **65**, 479–487 (2003).
21. L. Grosenick, T. S. Clement, R. D. Fernald, Fish can infer social rank by observation alone. *Nature* **445**, 429–432 (2007).
22. 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).
23. R. P. Gazes, N. W. Chee, R. R. Hampton, Cognitive mechanisms for transitive inference performance in rhesus monkeys: Measuring the influence of associative strength and inferred order. *J. Exp. Psychol. Anim. Behav. Process.* **38**, 331–345 (2012).
24. G. Jensen, H. S. Terrace, V. P. Ferrera, Discovering implied serial order through model-free and model-based learning. *Front. Neurosci.* **13**, 878 (2019).
25. C. J. C. H. Watkins, P. Dayan, Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
26. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction, Adaptive Computation and Machine Learning* (MIT Press, Cambridge, MA, ed. 2, 2018).
27. R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis* (Wiley, New York, 1959).
28. B. Carpenter *et al.*, Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
29. Stan Development Team, *Stan Modeling Language User's Guide and Reference Manual*, Version 2.28. https://mc-stan.org (2021).
30. 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).
31. O. Jensen, J. E. Lisman, Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends Neurosci.* **28**, 67–72 (2005).
32. S. Majerus, Verbal working memory and the phonological buffer: The question of serial order. *Cortex* **112**, 122–133 (2019).
33. S. Chen, K. B. Swartz, H. S. Terrace, Knowledge of the ordinal position of list items in rhesus monkeys. *Psychol. Sci.* **8**, 80–86 (1997).
34. 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).
35. G. Jensen, V. P. Ferrera, H. S. Terrace, Positional inference in rhesus macaques. *Anim. Cogn.* **25**, 73–93 (2021).
36. P. Dayan, "Exploration from generalization mediated by multiple controllers" in *Intrinsically Motivated Learning in Natural and Artificial Systems*, G. Baldassarre, M. Mirolli, Eds. (Springer, 2013), pp. 73–91.
37. E. Schulz *et al.*, Structured, uncertainty-driven exploration in real-world consumer choice. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 13903–13908 (2019).
38. C. M. Wu, E. Schulz, M. Speekenbrink, J. D. Nelson, B. Meder, Generalization guides human exploration in vast decision spaces. *Nat. Hum. Behav.* **2**, 915–924 (2018).
39. L. Lehnert, M. L. Littman, M. J. Frank, Reward-predictive representations generalize across tasks in reinforcement learning. *PLOS Comput. Biol.* **16**, e1008317 (2020).
40. I. Levy, J. Snell, A. J. Nelson, A. Rustichini, P. W. Glimcher, Neural representation of subjective value under risk and ambiguity. *J. Neurophysiol.* **103**, 1036–1047 (2010).
41. H. Pushkarskaya, M. Smithson, J. E. Joseph, C. Corbly, I. Levy, Neural correlates of decision-making under ambiguity and conflict. *Front. Behav. Neurosci.* **9**, 325 (2015).
42. B. Y. Hayden, S. R. Heilbronner, M. L. Platt, Ambiguity aversion in rhesus macaques. *Front. Neurosci.* **4**, 166 (2010).
43. A. Shenhav, R. L. Buckner, Neural correlates of dueling affective reactions to win-win choices. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10978–10983 (2014).
44. B. Malinowski, *Magic, Science and Religion, and Other Essays* (Beacon Press, Boston, 1948).
45. J. L. Risen, Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychol. Rev.* **123**, 182–207 (2016).
46. J. M. Pearce, G. Hall, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
47. C. D. Grossman, B. A. Bari, J. Y. Cohen, Serotonin neurons modulate learning rate through uncertainty. *Curr. Biol.* **32**, 586–599.e7 (2022).
48. K. Iigaya, S. Yi, I. A. Wahle, K. Tanwisuth, J. P. O'Doherty, Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nat. Hum. Behav.* **5**, 743–755 (2021).
49. A. Lak, W. R. Stauffer, W. Schultz, Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2343–2348 (2014).
50. L. K. Son, R. Sethi, Metacognitive control and optimal learning. *Cogn. Sci.* **30**, 759–774 (2006).
51. D. H. Brainard, The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
52. R. A. Rescorla, A. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement" in *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokasy, Eds. (Appleton-Century-Crofts, 1972), pp. 64–99.
53. Y. H. Jin, G. Jensen, J. Gottlieb, V. Ferrera, Superstitious learning of abstract order from random reinforcement, Figshare. https://figshare.com/articles/dataset/Supplementary_data_package_zip/19175612. Deposited 15 February 2022.