

The Role of Memory in Counterfactual Valuation

Natalie Biderman¹, Samuel J. Gershman², and Daphna Shohamy^{1, 3}

¹Department of Psychology, Mortimer B. Zuckerman Mind, Brain, Behavior Institute, Columbia University

²Department of Psychology and Center for Brain Science, Harvard University


³The Kavli Institute for Brain Science, Columbia University

Value-based decisions are often guided by past experience. If a choice led to a good outcome, we are more likely to repeat it. This basic idea is well-captured by reinforcement-learning models. However, open questions remain about how we assign value to options we did not choose and which we therefore never had the chance to learn about directly. One solution to this problem is proposed by policy gradient reinforcement-learning models; these do not require direct learning of value, instead optimizing choices according to a behavioral policy. For example, a logistic policy predicts that if a chosen option was rewarded, the unchosen option would be deemed less desirable. Here, we test the relevance of these models to human behavior and explore the role of memory in this phenomenon. We hypothesize that a policy may emerge from an associative memory trace formed during deliberation between choice options. In a preregistered study ($n = 315$) we show that people tend to invert the value of unchosen options relative to the outcome of chosen options, a phenomenon we term *inverse decision bias*. The inverse decision bias is correlated with memory for the association between choice options; moreover, it is reduced when memory formation is experimentally interfered with. Finally, we present a new memory-based policy gradient model that predicts both the inverse decision bias and its dependence on memory. Our findings point to a significant role of associative memory in valuation of unchosen options and introduce a new perspective on the interaction between decision-making, memory, and counterfactual reasoning.

Keywords: decision-making, memory, counterfactual reasoning, reinforcement learning, policy gradient models

Supplemental materials: <https://doi.org/10.1037/xge0001364.supp>

This article was published Online First May 18, 2023.

Natalie Biderman  <https://orcid.org/0000-0001-6691-4188>

Data and analysis codes are publicly available on Github and can be accessed at https://github.com/NatalieBiderman/The_role_of_memory_in_counterfactual_valuation. The study design, hypotheses, and analysis plan were preregistered on Open Science Framework prior to data collection and can be accessed at <https://osf.io/qad57/>.

The authors thank Anastasia Shuster for sparking the idea of the memory manipulation. They also thank Dan Biderman for insightful discussions and for providing feedback on an earlier draft. They are grateful for funding support from the NSF (1822619), NIMH/NIH (MH121093), the Templeton Foundation (60844), and the Multi-University Research Initiative Grant (ONR/DoD N00014-17-1-2961).

The authors have no conflicting interests to disclose.

Natalie Biderman served as lead for data curation, formal analysis, writing—original draft. Samuel Joseph Gershman served as lead for modeling and served in a supporting role for formal analysis, funding acquisition, writing—original draft. Daphna Shohamy served as lead for funding acquisition, resources, supervision, writing—review and editing and served in a supporting role for writing—original draft. Natalie Biderman, Samuel Joseph Gershman, and Daphna Shohamy contributed to conceptualization equally. Natalie Biderman and Samuel Joseph Gershman contributed to writing—review and editing equally.

Correspondence concerning this article should be addressed to Natalie Biderman or Daphna Shohamy, Department of Psychology, Mortimer B. Zuckerman Mind, Brain, Behavior Institute, Columbia University, 3277 Broadway, New York, NY 10025, United States. Email: natalie.biderman@columbia.edu or ds2619@columbia.edu

Memory plays a crucial role in our lives, allowing past experiences to inform future decisions. One way in which the past experience shapes decisions is through the learning of habits. Repeated experience with a positive outcome leads to the development of a behavioral tendency to repeat the same decision in the future (Balleine, 2018; O'Doherty et al., 2017). However, for many decisions, one has to choose between options that were never directly experienced before and for which the outcome is not known. For example, consider all the options that were never chosen: how do we learn to assign value to unchosen options, for which the possible outcome remains unknown?

Previous research in humans and animals has shown that inferences about value can extend beyond the particular experience of the organism. For example, motivational states can change the value of items (Pompilio et al., 2006; Pompilio & Kacelnik, 2005) and value information can be generalized based on perceptual features (Ghirlanda & Enquist, 2003; Kahnt et al., 2012; Kahnt & Tobler, 2016), or based on learned associations formed during encoding (Biderman et al., 2020; Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012). Here we propose that associative memory processes that are engaged during deliberation between choice options shape the valuation of unchosen options for which value is not learned explicitly.

In a recent study, we had people make decisions between pairs of options and learn the outcomes of their chosen options after a delay. We found that after learning the value of the chosen option, people updated the value of the unchosen option in the opposite

direction (Biderman & Shohamy, 2021). We refer to this effect as an “inverse decision bias.” When the chosen option proved to be valuable, the unchosen option was deemed less valuable; when the chosen option disappointed, the unchosen option became more attractive (for a related effect, see Marciano-Romm et al., 2016).

This inverse relationship resonates with previous studies showing that people continue to separate the value of choice options after the decision was made (Bem, 1967; Festinger, 1957; Henkel & Mather, 2007; Lind et al., 2017; Mather et al., 2000; Palminteri et al., 2015, 2017; Sharot et al., 2010; Voigt et al., 2017). However, while in the real world there is often a delay between choices and outcomes, none of these previous studies tested how delayed feedback for the chosen option affects the inferred value of the unchosen option. In such cases, we predict that associative memory plays a significant role. We suggest that deliberation binds choice options to each other in memory before any learning takes place, such that subsequent value information for one option affects the value of the other option (Liashenko et al., 2020; Wimmer & Shohamy, 2012; Wong et al., 2019). Indeed, in previous work, we found across five different datasets that the extent of the inverse decision bias was correlated with how well people remembered the pairs they deliberated about (Biderman & Shohamy, 2021).

A key goal of the present study was to advance understanding of the relationship between memory and the inverse bias in decisions. The previous results showed a replicable correlation between memory and decisions, but could not show whether memory was driving the inverse decision bias. It could be that memory for choice pairs is affected by the inverse inference of value, rather than causing it, or there could be a third variable that explains both the memory and the value inference effects.

To address these possibilities, in the current study we introduced an experimental manipulation that specifically targeted memory, and measured what effect, if any, this manipulation of memory had on choice. We reasoned that there are two general possibilities: First, it is possible that memory does not drive the inverse decision bias. If so, we would expect that a manipulation of memory will not affect the inverse decision bias. Second, it is possible that the inverse decision bias is memory-driven, that is, to revise the value of the unchosen option in relation to the chosen option, people *need* to remember that those two options appeared together. If so, we would expect that a manipulation of memory will lead to changes in the inverse decision bias.

To experimentally differentiate between these predictions, we used a well-known memory manipulation: retroactive interference (Jenkins & Dallenbach, 1924; McGeoch, 1932; Sosic-Vasic et al., 2018), in which associative pairs are recombined with novel items, thereby weakening the original memory trace. We integrated this manipulation into our experimental design of value-based decision-making, allowing us to compare, within participants, the effect of this memory manipulation on inverse bias. Specifically, in a large-scale preregistered study ($n = 315$), we had participants make decisions between pairs of items, and before learning the outcomes of their choices, we introduced retroactive interference for half of the pairs (by pairing the chosen item with a novel item). This allowed us to test whether inducing a memory-weakening manipulation before any value learning took place would alter the inverse decision bias for the manipulated pairs (henceforth, Interference pairs) compared to the remaining pairs (henceforth, Baseline pairs).

An additional goal of the current study is to provide a computational framework for the role of memory in counterfactual value assignment.

The inverse decision bias has significant implications for theories of value updating, and specifically for reinforcement learning (RL) models (Sutton & Barto, 1998). In value-based RL models, such as Q-learning (Watkins & Dayan, 1992), an agent learns the value of a set of discrete actions through direct experience and maps these estimated values onto a behavioral policy. A decision is made by comparing action values of choice options. Importantly, value updating takes place only for chosen options for which outcomes were directly experienced. Accordingly, standard Q-learning cannot account for the inverse decision bias, which involves updating the values of unchosen options that were never experienced (see Palminteri et al., 2015, 2017, for models that also update the value of unchosen options when their hypothetical outcomes are revealed).

Interestingly, there is another class of RL models—policy gradient models (Williams, 1992)—that explicitly predict the inverse decision bias. Policy gradient methods are behind some of the major advances in robotics (Peters & Schaal, 2006) and artificial intelligence more generally (Graves et al., 2016). They have also attracted attention in cognitive neuroscience (Bennett et al., 2021; Li & Daw, 2011; also in the form of actor-critic models, Joel et al., 2002; O’Doherty, 2004). In policy gradient models, the focus is on learning the parameters of a policy function rather than learning the value of specific state-action pairs (Bennett et al., 2021; Dayan & Abbott, 2001; Li & Daw, 2011). As in Q-learning, the agent wants to maximize the expected reward. But here, the agent makes decisions according to a behavioral policy and uses observed outcomes to adjust the parameters of the policy (following the gradient of the average reward with respect to the policy parameters). Importantly, all of the policy’s parameters are updated regardless of the specific actions taken. For a binary choice between two options, a simple parametrization of the policy is a logistic function of the difference between scalar policy parameters (“choice propensities”) of each option. For this parametrization, the signature of a policy gradient model is that the outcome of a chosen option will be used to update the choice propensities for both the chosen and the unchosen options (Dayan & Abbott, 2001). In particular, the propensities for the chosen and unchosen options will be updated in opposite directions (see “Method” section for mathematical description of the phenomenon), just as observed in our empirical data (Biderman & Shohamy, 2021).

While this kind of policy gradient model predicts the behavioral pattern of the inverse decision bias, it does not explain how the bias is affected by associative memory. In most value-based decision-making tasks (e.g., bandit tasks), a limited number of choice options repeat multiple times so that they are mostly remembered very well. Outside the lab, however, people clearly do not remember all their decisions equally, nor are they likely to manifest an equal amount of inverse decision bias following each decision.

In the current study, we devised a new memory-based policy gradient model that predicts both the inverse decision bias and its modulation by associative memory. We used the classical policy gradient model and added a memory component that controls the extent of value update for the unchosen option, based on how well people remembered which chosen option was paired with it. We named this model the variable-memory model since the memory component was allowed to vary across the different choice pairs. We compared this model to two other models. The first is a fixed-memory

model, where value update for the unchosen option was modulated by the overall memory (by taking the average memory scores across all choice pairs). The fixed-memory model takes into account general memory performance but does not predict value inference based on memory for specific choice pairs. The second model is a perfect-memory model, wherein value update was not modulated by memory at all (i.e., the value of the unchosen option is always updated). This model is essentially the classical policy gradient model that assumes perfect memory for choice pairs. We predicted that only the variable-memory model would capture behavioral differences between the Interference and Baseline pairs.

In summary, the contributions of the current study are to explore the role of memory in counterfactual value assignment, in two ways. First, we go beyond the previous memory and inverse-bias correlation finding and experimentally manipulate the strength of associative memory to test how it modulates the inverse decision bias. Second, we present a new memory-based policy gradient model that predicts the relationship between inverse decision bias and associative memory. The experiment presented here is a large-scale replication ($n = 315$) of a pilot study ($n = 108$), and the experiment, predictions, and analysis plan were all preregistered on Open Science Framework (<https://osf.io/qad57/>).

Method

Preregistration and Sample Size Determination

The experiment was preregistered on Open Science Framework (<https://osf.io/qad57/>; Biderman et al., 2022). Prior to data collection, we ran a Pilot study ($n = 108$) which confirmed our main hypotheses and guided later sample size decisions. Specifically, we confirmed that our memory manipulation was successful at modulating pairs memory and inverse decision bias (see Supplementary Tables 1–3 in the online supplemental materials). The goal of the current study was to replicate these findings using a substantially larger sample size. To determine the sample size, we performed a power analysis aimed at reaching above 99% power. We used the Pilot study to compute an effect size for our main behavior prediction. Namely, the comparison between the tendency to select $S_{\text{unchosen}} = 0$ items (unchosen items previously paired with unrewarded chosen items) in Interference versus Baseline trials in the Final Decisions phase (see Procedure). To this end, we ran a paired t -test to estimate the statistical difference between the two conditions, $t(107) = 2.57$, $p = .011$, Cohen's $d = 0.25$. Using this effect size as a point of reference, we found that 302 participants would give us 99% power to detect a difference in a new sample with a significance level of .05. We therefore defined our sample size to include 305 participants that do not meet any of our predefined exclusion criteria (see below). We ran more participants expecting some to be excluded, and ended up having 315 participants who passed the exclusion criteria. The final sample size was therefore 315 participants. Our main findings remain substantial when including only 305 participants. Notably, 315 participants are more than 2.5 the size of the pilot study, the recommended ratio for replication studies (Simonsohn, 2015).

Participants

The research was approved by the Institutional Review Board (IRB) at Columbia University through Columbia IRB Protocol #AAAI1488. 315 Mechanical Turk participants took part in the

study for an average payment of \$10.2 (M_{age} : 29.1 ± 4.3 , 173 female, 138 male, 4 other). All participants provided informed consent for their participation in the experiment. One hundred and eight participants participated in the Pilot study (M_{age} : 29.6 ± 3.7 , 54 female, 54 male). Additional 25 and 9 participants from the current and the Pilot studies, respectively, were removed from analyses because they met our predefined exclusion criteria for online studies (see below, all exclusion criteria were preregistered on OSF: <https://osf.io/qad57/>). To avoid age- and location-related confounds, we added a restriction on Mechanical Turk to include only U.S.-based participants within the age range of 18–36. We also restricted participants to an approval rate of above 90%.

Exclusion Criteria

We applied the following exclusion criteria which were all aimed to ensure that the online participants were attending the task: (1) below chance performance (probability to choose rewarded items below .5) in the Final Decisions phase for chosen pairs, indicating participants who did not learn the new values of chosen paintings (the exclusion performance was computed across conditions); (2) more than 25 missed responses in either the Outcome Learning phase (where participants had to register the outcomes they observed) or the Final Decisions phase; (3) more than 25 events where participants were browsing a different window in any experimental phase (blur-focus events detected using jsPsych library, de Leeuw, 2015); (4) more than 10 trials in the Deliberation phase where responses were too fast (below 300 ms; these trials were accompanied with a warning), signifying no actual deliberation; and (5) more than 10 failed attempts to answer a comprehension quiz administered after instructions in any experimental phase. Participants who met at least one of these exclusion criteria were removed from all analyses.

Materials

Stimuli were images of representational paintings depicting people. The stimuli were collected by Celia Durkin from various online databases, and a subset of these stimuli are published and available online (Durkin et al., 2020). The stimuli were converted to 300×300 pixel size and were presented on a gray background (RGB: 128, 128, 128; Figure 1).

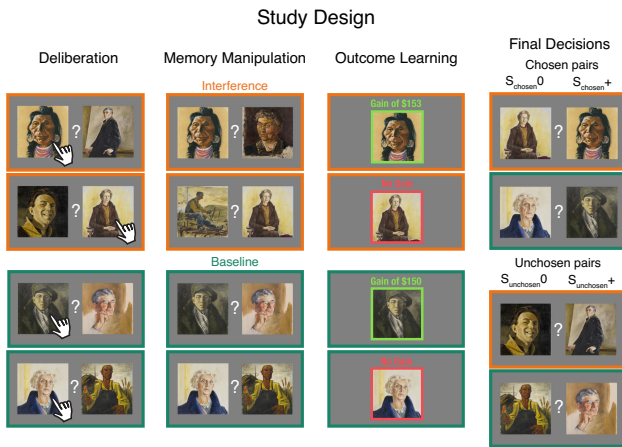
Procedure

We modified the paradigm used previously to demonstrate the inverse decision bias (Biderman & Shohamy, 2021; Figure 1). The study included the following consecutive phases: (1) Pretask Ratings, (2) Deliberation, (3) Memory Manipulation, (4) Outcome Learning, (5) Final Decisions, (6) Surprise Memory, and (7) Outcome Evaluation. Each phase began with instructions followed by a comprehension quiz. Participants were not informed about the upcoming phases. On average, the experiment took 56 ± 0.84 min.

Pretask Ratings

Participants were presented with 60 paintings and were asked to rate their liking of each painting using a continuous scale (from “not at all” to “very much,” responses were then scaled from 0 to

Figure 1
Study Design



Note. Participants act as art dealers choosing between pairs of paintings which are later sold in two separate auctions. In the *Deliberation* phase, participants make a series of choices between pairs of paintings. Next, in the *Memory Manipulation* phase, participants engage in another auction. Unbeknownst to them, the pairs in this second auction are divided into two groups: Half of the pairs were previously chosen items coupled with novel items (*Interference pairs*, marked here with an orange frame for illustration purposes only), while the other half were the same as in the first auction (*Baseline pairs*, shown here in a turquoise frame). Next, in the *Outcome Learning* phase, chosen paintings are presented one at a time with the profit they made in the auction: either “gain” (green frame with dollar amount in gain) or “no gain” (red frame). The *Final Decisions* phase aimed to measure biases in choices: participants are asked to choose the most valuable between two options. Unbeknownst to them, the pairs are either made up of two previously chosen paintings (rewarded and unrewarded in the auction, denoted as $S_{\text{chosen}+}$ and $S_{\text{chosen}0}$, respectively) or two previously unchosen paintings (initially presented with $S_{\text{chosen}+}$ and $S_{\text{chosen}0}$, denoted as $S_{\text{unchosen}+}$ and $S_{\text{unchosen}0}$, respectively). The two pair types are randomly intermixed within the same experimental condition (Interference or Baseline). The study additionally included three phases not shown here: A *Pretask Rating phase*, wherein participants report their liking of each painting prior to the experiment to control for any baseline preferences among the options. A *Surprise Memory phase*, after the Final Decision phase, wherein participants are tested for associative memory of the deliberation pairs. And finally, an *Outcome Evaluation phase*, wherein participants are asked to estimate the auction profit of all paintings from the Deliberation phase, including those they did not choose, as well as novel paintings from the Memory Manipulation phase. In all phases, trials from both experimental conditions (Interference or Baseline) were randomly intermixed. Stimuli were art images depicting people (Durkin et al., 2020). The figure includes an example of a few images, all by Clyfford Still, courtesy of the Clyfford Still Museum, Denver, CO © City and County of Denver (see Supplementary Figure 1 for a full description of the images presented in the current figure). Images in the actual experiment were by a variety of painters, randomly intermixed. If viewing in grayscale, orange equals light gray, and turquoise equals dark gray. See the online article for the color version of this figure.

100). At the end of the phase, the paintings were sorted by their ratings, and the 40 paintings rated in the middle of the distribution were selected to serve as the experimental stimuli in the study. Thirty-two of the paintings were randomly distributed across 16 Deliberation pairs, and 8 paintings were designated as novel paintings in the Memory Manipulation phase (see below).

Deliberation

After completing the ratings, participants were instructed to act as art dealers. They were told that two separate auctions are about to take place, and for each they will deliberate between pairs of paintings and decide which painting would be more profitable in the upcoming auctions. They were also told that they would only see the outcomes from one of the auctions. During the Deliberation phase (Auction #1), participants were shown 16 pairs of paintings and were asked to deliberate and choose one of the paintings in each pair. Unbeknownst to them, half of the pairs were designated to undergo an interference manipulation later on (see below, Interference pairs), and the other half were designated not to undergo interference (Baseline pairs), their order was randomly intermixed. The two pair types did not differ in the first auction and they were not physically marked in any way throughout the experiment. Participants were instructed to take their time and were given up to 10 s to make a decision. If they did not respond during this time, they were prompted to make the decision again and to respond more quickly. To ensure deliberation and encoding of the pairs, we told participants they will practice the decisions a couple of times before committing to their final choices. The 16 Deliberation trials were repeated three times in three separate blocks, with random order of trials within each block. In subsequent phases, we used the choices made in the last Deliberation block. To increase deliberation, participants were asked to write down the reasons for their decisions using a text box. The text boxes appeared once for every deliberation pair across the first two blocks. Participants were told that one of their decisions will be played out for real, and they will receive 1% of their chosen painting’s auction earnings, if indeed it resulted in a gain. At the end of the Outcome Learning phase (see below), all participants were informed that they received extra \$1.5 bonus money (we made sure one of the chosen items gained \$150).

Memory Manipulation

After the first auction, participants moved on to the second auction. Unbeknownst to them, the purpose of the second auction was to manipulate the associative memory of pairs from the first auction (those presented in the Deliberation phase).

The decision pairs in the second auction were divided into two groups. Half of the pairs (Interference trials) were constructed from novel paintings coupled with paintings that were chosen from the first auction, out of the Interference pairs. The position of the novel painting in each pair was counterbalanced across pairs. This new coupling introduced retroactive interference (Jenkins & Dallenbach, 1924; McGeoch, 1932; Sosis-Vasic et al., 2018), intended to weaken the association created in the Deliberation phase. The other half of pairs (Baseline trials) were the Baseline pairs from the first auction, with no change. Because we were interested in the effect of retroactive interference on half of the pairs from the first auction, we had to make sure that the second auction did not override the memory of the other, unmanipulated, Baseline pairs. We therefore repeated the Baseline pairs in the second auction.

As in the Deliberation phase, all pairs repeated three times and participants were told they could practice these decisions a couple of times before committing to their final choices and they were asked to explain their decisions using a text box.

Outcome Learning

After completing the second auction, participants were told they will learn the outcomes of the first auction (i.e., for decisions made in the Deliberation phase). Notably, choice outcomes were presented in a separate phase rather than after every decision was made. This was meant to tease apart the effects of working memory on value updating from long-term associative memory, the focus of our study. We assumed that after making a choice, participants could still maintain both choice options in their working memory and feedback for the chosen option could transfer to the unchosen option just by virtue of being activated simultaneously in working memory. Thus, at the end of the second auction, we presented the first auction outcomes for chosen paintings alone. For every Deliberation pair type (Baseline, Interference), half of the chosen items were randomly assigned as rewarded items (denoted as S+), with earnings centered around \$150 and *SD* of \$10. The other half were assigned as unrewarded items, with \$0 earnings (denoted as S0). Paintings were presented in the center of the screen alongside a colored outcome and frame (green for S+, red for S0, Figure 1) for 2 s. To facilitate learning, we repeated each painting six times across three blocks, in a randomized order. To ensure participants' attention, we asked them to press the spacebar to see the outcome auction and then to register the outcome by pressing a corresponding key while it is presented on the screen (up arrow for gain, down arrow for no gain). If they missed an outcome registration, they saw a warning asking them to respond faster.

Final Decisions

This phase sought to measure the inverse decision bias. We asked participants to create a portfolio of high-valued paintings. To this end, they had to make a new series of decisions between pairs of paintings, choosing the most valuable painting in each pair. To incentivize their performance, participants were told they could earn extra bonus money based upon their performance in this phase. Each trial in this phase presented either two previously chosen stimuli from the first auction (rewarded and unrewarded, denoted as $S_{\text{chosen}+}$ and $S_{\text{chosen}0}$, respectively), or two previously unchosen stimuli from the first auction (denoted as $S_{\text{unchosen}+}$ and $S_{\text{unchosen}0}$, presented earlier with $S_{\text{chosen}+}$ and $S_{\text{chosen}0}$, respectively, Figure 1). Importantly, the chosen and unchosen pairs were constructed from either two Baseline or two Interference items, so we could assess choices in every experimental condition separately and compare performance between the two conditions. This yielded 16 unique chosen pairs and 16 unique unchosen pairs, in every experimental condition, for a total of 72 unique pairs across conditions. To increase the number of trials, we repeated the decision trials four times across four separate blocks, each including all 72 pairs, randomly intermixed. In each block, the rewarded painting appeared on the left and the right sides an equal number of times. Participants had 2.5 s to make a decision and, if they failed to respond during this time, they were shown a warning asking them to respond faster. Unbeknownst to participants, the potential bonus money was based only on their performance for the chosen pairs and was up to \$2 (average bonus money was 1.69 ± 0.61).

Surprise Memory

To test for associative memory of the Deliberation pairs (pairs from the first auction), we presented 32 pairs of paintings and

asked participants whether each pair was intact or recombined. Intact pairs were pairs of paintings that appeared in the first auction. Recombined pairs were pairs that included a chosen painting and an unchosen painting from the first auction that did not previously appear together, and the combination was made within the same experimental condition (e.g., chosen and unchosen items from Interference pairs were coupled together). If participants responded "intact," they were also asked to indicate which of the two paintings they previously had chosen in the Deliberation phase.

Outcome Evaluation

In this final phase, we told participants that all paintings were sent to auction, including those they did not choose, and asked them to estimate the auction outcomes. We presented previously chosen and unchosen paintings from the Deliberation phase and the novel paintings from the Memory Manipulation phase. Paintings were presented in a random order and for each painting we asked participants whether the painting resulted in a gain or not, and then to rate how confident they are in their response from 1 ("completely unsure") to 6 ("completely sure").

At the end of the experiment, we asked participants about their decision strategies throughout the task. We were specifically interested to examine the possibility that participants might have adopted an inverse heuristic. That is, they might have assumed that for every decision made in the Deliberation phase, one painting would result in a gain and the other would not. To this end, we first asked participants what was their decision strategy in the Deliberation phase and then we asked them how they decided between pairs of paintings for which they received direct feedback (chosen pairs) and for which they received no feedback (unchosen pairs) in the Final Decisions phase (see analysis of the inverse heuristic in Supplementary Text 3 in the online supplemental materials).

Data Analysis

Data analysis followed the same rationale as in Biderman and Shohamy (2021) with the addition of another predictor for condition type (Baseline vs. Interference) when relevant.

Data were preprocessed in Python and analyzed in R. All results were analyzed with Bayesian generalized linear models using the "rstanarm" package (Goodrich et al., 2022), which performs approximate Bayesian inference over the regression coefficients. Instead of a maximum-likelihood procedure that provides a single point-estimate for each coefficient, Bayesian inference targets the full posterior distribution of each coefficient, which combines our prior assumptions and the observed data. We approximated the posterior distribution of the models' coefficients using Markov Chain Monte Carlo methods, and improved the fidelity of our samples by running multiple independent sampling processes ("chains"). For every model, we used default priors and ran six chains with 4,000 iterations each (2,000 iterations per chain were used as warm-up). To determine convergence of each chain, we made sure that for all model coefficients, the effective sample size of simulation draws was greater than 900 and the R-hat statistic was around 1.0 (Gelman et al., 2013; Gelman & Rubin, 1992). To evaluate our effects, for each regression coefficient of interest, we report the median of the posterior samples and their 95% highest density interval (HDI). Since a regression coefficient of zero indicates no relationship between a

predictor and an outcome, we determined that a variable reliably predicts an outcome only if its 95% HDI excluded zero.

Moreover, when possible, we ran multilevel (or hierarchical) models, in which a participant's regression coefficient is drawn from a group-level coefficient distribution. Such an approach is more robust to outlier subjects and observations. The parameters of this group-level distribution are of special importance, since they indicate whether an effect is present or not across all subjects. In all multilevel models, all predictors varied by participants.

Analysis of Memory Performance

To test for differences in associative memory between our experimental conditions, we ran a multilevel Bayesian logistic regression that predicted the probability of an accurate response in the Surprise Memory phase as a function of the experimental condition (Baseline or Interference, centered predictor). Accurate responses are those wherein participants correctly responded "intact" when the pair was intact, and "recombined" when it was recombined. The coefficient of interest is the condition coefficient. To detect differences between conditions in signal detection theory measures (d-prime, number of hits, number of false alarms), we computed the measures using the "psycho" R package (Makowski, 2018), for each participant in each condition, and ran a Bayesian linear regression predicting the specific measure as a function of condition. Signal detection theory analyses were exploratory and were not preregistered (<https://osf.io/qad57/>).

Analysis of Choices in the Final Decisions Phase

We ran a multilevel Bayesian logistic regression that predicted the probability of choosing S+ over S0 as a function of choice (Chosen or Unchosen pairs, centered predictor), condition (Baseline or Interference, centered predictor), the difference in pretask liking ratings between the items in each pair (S+ minus S0), and all possible interactions between the predictors ($p(\text{select S+}) = \beta_0 + \beta_1\text{choice} + \beta_2\text{condition} + \beta_3\Delta\text{ratings} + \beta_4\text{choice:condition} + \beta_5\text{choice}:\Delta\text{ratings} + \beta_6\text{condition}:\Delta\text{ratings} + \beta_7\text{choice:condition}:\Delta\text{ratings}$). We normalized each participant's ratings to control for differences in the overall use of the scale using a z-score normalization (i.e., some subjects tend to use higher/lower values of the continuous scale), and then subtracted the normalized rating of the unrewarded painting (S0) from the rewarded painting (S+) in each decision trial. Trials where participants missed a response were excluded from analyses.

Our primary measure of interest is the difference between Baseline and Interference conditions in unchosen pairs. Following the rationale in Biderman and Shohamy (2021) we rearranged the model equation, effectively decomposing the model into four regression models, creating all possible combinations of choice (Chosen, Unchosen) and condition (Baseline, Interference). Each such model is predicting the probability of gain choice as a function of ratings. We then computed a separate coefficient and a separate intercept term for our four regression models. The coefficient is a slope term indicating the influence of ratings on choice (computed as $\beta_3 + \beta_5*\text{choice} + \beta_6*\text{condition} + \beta_7*\text{choice}*\text{condition}$, e.g., for chosen-baseline condition, choice = 1, condition = -1, hence the slope term equals $\beta_3 + \beta_5 - \beta_6 - \beta_7$). The intercept term quantifies the tendency to choose rewarded items when there is no difference in ratings (computed as $\beta_0 + \beta_1*\text{choice} + \beta_2*\text{condition} +$

$\beta_4*\text{choice}*\text{condition}$, for example, for chosen-baseline condition, the intercept term is $\beta_0 + \beta_1 - \beta_2 - \beta_4$). Our measure of interest is the difference between the intercept terms of the Baseline and Interference conditions for unchosen pairs.

Analysis of Reaction Times in the Final Decisions Phase

The focus of this analysis is on how choices affect reaction times in unchosen pairs. To minimize the number of predictors, we included only unchosen pairs in this analysis (also note that for chosen pairs most choices are of S+, so there is not enough variability in responses to compute reliable estimates). We ran a Bayesian generalized linear mixed model, predicting raw reaction times as a function of choice (S+ or S0, centered predictor), condition (Baseline or Interference, centered predictor), and their interaction (reaction times = $\beta_0 + \beta_1\text{choice} + \beta_2\text{condition} + \beta_3\text{choice:condition}$). Following Lo and Andrews (2015), we used a Gamma distribution to model the dependent measure with a logarithmic link function. Using the `brm()` function in the "brms" R package (Bürkner, 2017), we implemented a default prior over the shape parameter (gamma distribution with a scale and shape parameters equal to 0.01), and a normal distribution centered at zero with 2 SDs, over the intercept and beta parameters. The prime measure of interest is the interaction coefficient which quantifies the difference between conditions in the effect of choice (choosing S+ vs. S0) on reaction times. To get a separate estimate of the effect of choice on reaction time for every condition, we plugged in the condition to the following term $\beta = \beta_1 + \beta_3*\text{condition}$.

Analysis of the Relationship Between Associative Memory and Inverse Decision

For every participant in each experimental condition, we computed two measures. The first was average accuracy in pairs memory responses collected during the Surprise Memory phase. The second was a decision bias score, measuring the contrast in decision tendency between chosen and unchosen pairs. This was operationalized as the mean probability to choose a rewarded item in chosen pairs minus unchosen pairs. Values closer to 1 signify a large inverse decision bias effect. We used a Bayesian linear regression to predict inverse decision bias as a function of memory accuracy and condition (Baseline and Interference, centered predictor; inverse decision bias = $\beta_0 + \beta_1\text{memory} + \beta_2\text{condition} + \beta_3\text{memory:condition}$).

We rearranged the model coefficients to get a memory term for every experimental condition (e.g., for Interference condition, condition = 1 and the slope term equals $\beta = \beta_1 + \beta_3$). The difference between the conditions was measured by the memory accuracy and condition interaction coefficient.

Model Description

We assume that the policy, that is, the probability of choosing an option, is given by a three-parameter logistic function:

$$\pi(\text{chosen}) = \frac{1}{1 + \exp[-\beta(\theta_{\text{chosen}} - \theta_{\text{unchosen}})]} \quad (1)$$

where θ_{chosen} and θ_{unchosen} are the "action propensities" for the soon-to-be-chosen and soon-to-be-unchosen actions, respectively. An action propensity is the tendency to select one option over the

other, and it is the difference between them that determines which option is selected. The difference is scaled by the inverse temperature parameter β which controls choice stochasticity. If the agent chooses their preferred action and observes a reward r , the model updates the policy parameters according to:

$$\Delta\theta = \alpha(r - V)\nabla_{\theta}\pi \quad (2)$$

where α is a learning rate, $V = E_{\pi}[V] = \pi(\text{chosen})V_{\text{chosen}} + \pi(\text{unchosen})V_{\text{unchosen}}$ is the policy-weighted average of the mean reward for each action, and $\nabla_{\theta}\pi$ is the gradient of the policy with respect to its propensity parameters. For the logistic policy of Equation (1), this gradient contains two partial derivatives:

$$\nabla_{\theta_{\text{chosen}}}\pi = \beta[1 - \pi(\text{chosen})] \quad (3)$$

$$\nabla_{\theta_{\text{unchosen}}}\pi = -\beta\pi(\text{chosen}) \quad (4)$$

The partial derivatives point in opposite directions, and therefore increasing the propensity for the chosen option will decrease the propensity of the unchosen option, and vice versa. This is a result of choosing the standard logistic policy parametrized by the difference between the two options' action propensities ($\theta_{\text{chosen}} - \theta_{\text{unchosen}}$). Such a policy is intuitive: it roughly amounts to the common practice of fitting a logistic regression model to subjects' binary choices based on the difference between their assigned values.

The model in its current formulation assumes perfect memory of the choice options. To account for imperfect memory, we let $p(\text{chosen}, \text{unchosen})$ denote a joint distribution of actions, representing the probability of recalling the chosen and unchosen options together. In order to match the experimental procedure from Biderman and Shohamy (2021), we consider the case where we observe the outcome for the chosen option and then update the unchosen option (hence we condition on the chosen option):

$$E[\nabla_{\theta_{\text{unchosen}}}\pi|\text{chosen}] = -\beta\pi(\text{chosen})[1 - \pi(\text{chosen})] / p(\text{unchosen}|\text{chosen}) \quad (5)$$

This equation is obtained by averaging the gradient in Equation (4) under the memory distribution (the conditional probability of remembering the unchosen option conditional on retrieving the chosen option). Accordingly, the inverse decision bias should increase with the conditional probability of recalling the unchosen option when presented with the chosen option, $p(\text{chosen}|\text{unchosen})$.

We compared three versions of the memory-based policy gradient model, which differed only in their assumptions about memory. In the *variable memory* model, we constructed a memory score for each item and used this score as a proxy for $p(\text{unchosen}|\text{chosen})$. Items in the Surprise Memory phase were presented twice, once within the actual pair that appeared in the Deliberation phase (intact pairs) and once within a recombined pair. For each item, we computed a memory score based on responses on these two trials. A score of 1 depicts a perfect memory, with two correct responses, a score of 0.5 depicts one correct response, and a score of 0 depicts no correct responses. In the *fixed memory* model, we used the same memory scores but replaced each item's score with the average across items. This model retains the average effect of memory on the inverse decision bias for each subject but eliminates all trial-by-trial variability. Thus, this model cannot explain the relationship between memory and decision variability within participants. Finally, the

perfect memory model assumes that $p(\text{unchosen}|\text{chosen}) = 1$ for all items. This model does not capture any variability in memory performance either within or between subjects.

Model Fitting

We used importance sampling with 9,000 samples to approximate the posterior distribution over parameters for each subject. The parameters consisted of the inverse temperature (β) and two learning rates (α_{chosen} and α_{unchosen}). We fit separate learning rates in order to attain greater modeling flexibility and to follow previous work testing the effects of chosen and unchosen options on learning and memory (e.g., Palminteri et al., 2015, 2017). The modeling results remain the same when fitting a single learning rate. The prior overall parameters were set to a uniform distribution (bounded between 0 and 15 for the inverse temperature, and between 0 and 1 for the learning rates). We used the prior distribution as the proposal distribution. The model was fit to the Deliberation choices only, so that the choices in the Final Decisions phase could be used as held-out data for model comparison. We submitted the log-likelihood for the Final Decisions' choices to a random-effects Bayesian model comparison (Rigoux et al., 2014).

Transparency and Openness

We report how we determined our sample size, all participant and data exclusions, all manipulations, and all measures in the study. The study design, hypotheses, and analysis plan were preregistered on Open Science Framework prior to data collection and can be accessed at <https://osf.io/qad57/>. Data and analysis codes have been made publicly available at Github and can be accessed at https://github.com/NatalieBiderman/The_role_of_memory_in_counterfactual_valuation. Study materials are available upon request. Data were analyzed using Python (Version 3.8.8, Van Rossum & Drake, 2000), using the packages *pandas* (Version 1.2.4, McKinney, 2010) and *NumPy* (Version 1.20.1, Harris et al., 2020), as well as R (Version 4.1.1, R Core Team, 2013), using the R-packages *ggplot2* (Version 3.3.5, Wickham, 2016), *tidyr* (Version 1.1.4, Wickham et al., 2023), *dplyr* (Version 1.0.7, Wickham et al., 2022), *rstanarm* (Version 2.21.1, Goodrich et al., 2022), *kableExtra* (Version 1.3.4, Zhu, 2021), *pwr* (Version 1.3.0, Champely, 2020), *lsr* (Version 0.5.1, Navarro, 2013), *cowplot* (Version 1.1.1, Wilke, 2020), *stringr* (Version 1.4.0, Wickham, 2019), *psycho* (Version 0.6.1, Makowski, 2018), and *brms* (Version 2.18.0, Bürkner, 2017).

Results

All results were analyzed with Bayesian generalized linear models (see details in the "Analysis" section). In each model, we estimated a posterior distribution for regression coefficients and reported the median and 95% HDI for coefficients of interest. If the 95% HDI of a coefficient excluded zero, we concluded that the corresponding variable reliably predicted the outcome.

Associative Memory Is Worse for Pairs That Underwent Interference

We first verified that the memory manipulation modulated associative memory of the Deliberation pairs. Indeed, accuracy in the Surprise Memory phase was lower for Interference pairs ($0.58 \pm$

0.01 [$M \pm SE$]) compared to Baseline pairs (0.66 ± 0.01 , coefficient in a model predicting accuracy by condition: $\beta_{\text{condition}} = -0.17$ [$-0.21, -0.12$]; Figure 2a). We ran an exploratory, unregistered, analysis to test whether the two conditions differed in their memory sensitivity. To this end, we computed d' -prime, a measure of sensitivity derived from signal detection theory (Macmillan & Douglas Creelman, 1991). Mirroring the accuracy results, d' -prime was higher in the Baseline condition ($d' = 0.85 \pm 0.05$) compared to the Interference condition ($d' = 0.45 \pm 0.04$; coefficient in a model predicting d' -prime by condition: $\beta_{\text{condition}} = -0.20$ [$-0.26, -0.14$]). The two conditions differed in their number of hits (Baseline: 5.80 ± 0.10 , Interference: 4.71 ± 0.11 ; $\beta_{\text{condition}} = -0.55$ [$-0.69, -0.40$]), but not in their number of false alarms (Baseline: 3.30 ± 0.11 , Interference: 3.35 ± 0.11 ; $\beta_{\text{condition}} = 0.03$ [$-0.13, 0.18$]).

Memory Interference Reduced the Inverse Inference of Value of Unchosen Items

Based on our theoretical framework, we hypothesized that weakening associative memory of the Deliberation pairs would reduce the inverse decision bias during the Final Decisions phase.

We found that for chosen pairs, participants consistently preferred $S_{\text{chosen}+}$ over $S_{\text{chosen}0}$, with no difference between the Interference and Baseline conditions (Interference: probability to select $S+$: 0.91 ± 0.01 , coefficient measuring the tendency to select $S+$: $\beta = 3.38$ [$3.17, 3.61$]; Baseline: 0.91 ± 0.01 , $\beta = 3.30$ [$3.10, 3.51$], coefficient measuring the difference between conditions: $\beta = 0.09$ [$-0.11, 0.28$], see “Analysis” section for model specification). This suggests that participants correctly learned the outcomes of their chosen items, and that associative memory strength did not modulate direct outcome learning.

Critically, our memory manipulation did modulate choices in unchosen pairs. In the Baseline condition, we replicated the inverse decision bias. Participants tended to select $S_{\text{unchosen}0}$ over $S_{\text{unchosen}+}$ (probability to select $S+$: 0.44 ± 0.01 , $\beta = -0.32$ [$-0.47, -0.18$], Figure 2b; for unchosen pairs, $S+$ items are paintings previously paired with a rewarded painting, and $S0$ items were previously paired with an unrewarded painting). Inverse decision bias was absent in the Interference condition. Participants were at chance-level when deciding between $S_{\text{unchosen}0}$ and $S_{\text{unchosen}+}$ (probability to select $S+$: 0.48 ± 0.01 , $\beta = -0.09$ [$-0.22, 0.06$]). The difference between conditions was substantial ($\beta = 0.24$ [$0.05, 0.41$]), suggesting that our memory manipulation modulated choices only in unchosen pairs, for which value was not directly learned and could only be inferred.

Interestingly, choice patterns in the Final Decisions phase were mirrored by reaction times. Previously we have shown that the tendency to select unchosen paintings previously paired with unrewarded paintings ($S_{\text{unchosen}0}$ over $S_{\text{unchosen}+}$) is accompanied with faster responses (Biderman & Shohamy, 2021). This pattern is expected if $S_{\text{unchosen}0}$ paintings are considered more valuable, as choices to gain rewards are faster than choices to avoid loss (Chen & Bargh, 1999; Duckworth et al., 2002). In the current study, we replicated this effect in the Baseline condition. Trials wherein $S_{\text{unchosen}0}$ items were selected were faster than trials wherein $S_{\text{unchosen}+}$ were selected (reaction times: 0.96 ± 0.01 , $S_{\text{unchosen}+}$: 1.01 ± 0.01 , coefficient quantifying the difference between $S_{\text{unchosen}0}$ and $S_{\text{unchosen}+}$: $\beta = 0.01$ [$0.01, 0.02$], see Analysis section for details). Importantly, this pattern was not replicated in the Interference condition. Reaction times were comparable across trials

wherein participants selected either $S_{\text{unchosen}0}$ or $S_{\text{unchosen}+}$ items ($S_{\text{unchosen}0}$: 0.97 ± 0.01 , $S_{\text{unchosen}+}$: 0.99 ± 0.01 , $\beta = 0.01$ [$-0.002, 0.01$]). The difference between conditions did not reach significance ($\beta = -0.01$ [$-0.02, 0.003$]).

Together, these results demonstrate that associative memory plays a significant role in the inverse inference of value (see Supplementary Text 1 in the online supplemental materials for analyses testing potential alternative explanations for the relationship between memory and inverse decision bias). Decreasing the association between choice options eliminates the inverse decision bias both in choices and reaction times (see Supplementary Text 2 and Supplementary Figure 2 in the online supplemental materials and Supplementary Table 4 for analysis of explicit evaluations of value mirroring the inverse decision bias). While the retroactive interference manipulation did not affect choices based on direct outcome learning (in chosen pairs) it did affect choices based on inference alone (in unchosen pairs), eradicating the tendency to update the value of unchosen options in opposition to the update for chosen options.

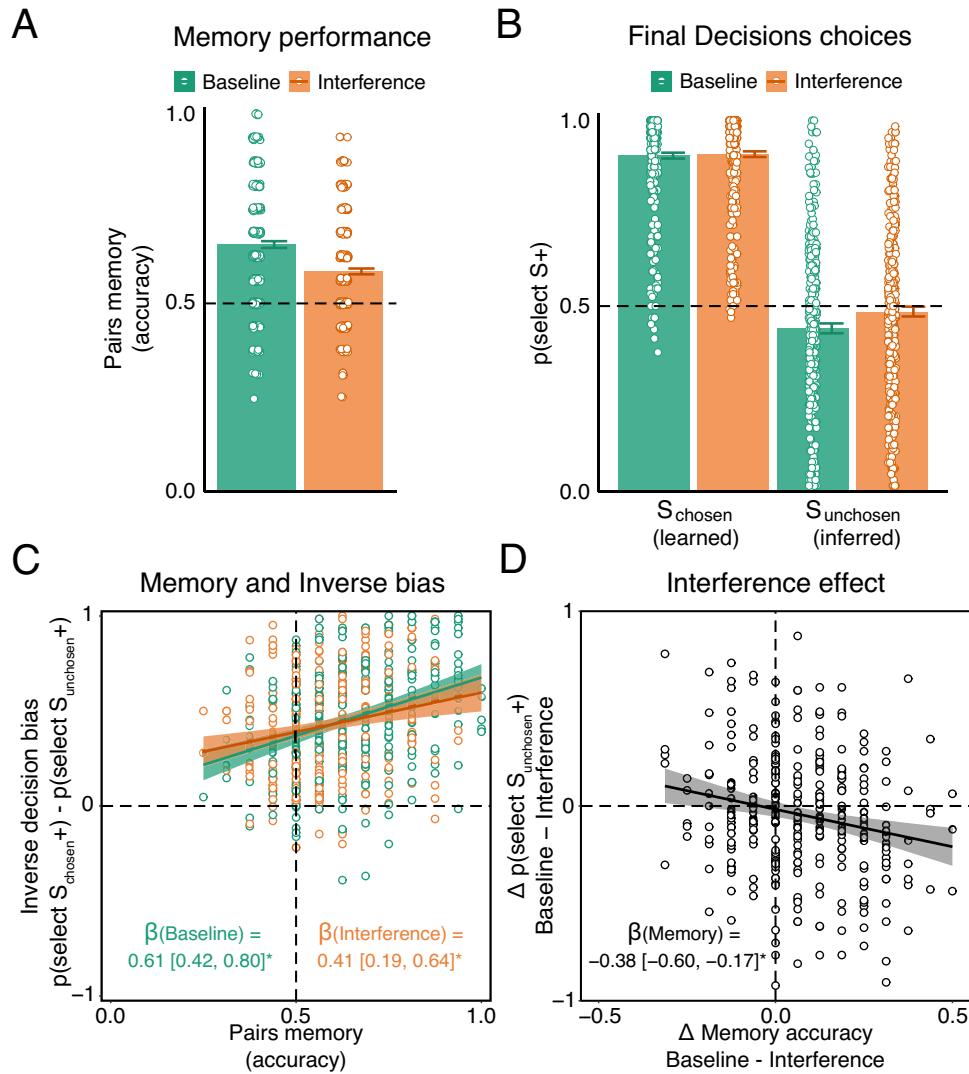
Associative Memory Predicts Decision Bias in Unchosen Pairs

In previous work, we found that individuals with better memory for the decision pairs also exhibited a larger decision bias effect (Biderman & Shohamy, 2021). Here, we test whether the extent of this correlation differs between the Baseline and Interference conditions. We found that for both conditions, memory predicted the inverse decision bias, Baseline: $\beta_{\text{memory}} = 0.61$ [$0.42, 0.80$], Interference: $\beta_{\text{memory}} = 0.41$ [$0.19, 0.64$] (Figure 2C). While the Baseline condition showed a larger effect descriptively, the difference between conditions did not reach significance, $\beta_{\text{memory:condition}} = -0.10$ [$-0.25, 0.05$]; see Analysis section for model specification. To assess whether it is the reduced memory in the Interference condition that predicts the reduced inverse decision bias in the same condition, we ran the following exploratory, unregistered analysis (see preregistration: <https://osf.io/qad57/>). For each participant, we computed difference scores between the Baseline and Interference conditions both in memory performance as well as in choices of unchosen pairs in the Final Decisions phase. We focused on unchosen pairs since the memory manipulation only affected unchosen trials (similar results are obtained if we compute the difference between inverse decision bias, that is, the tendency to select rewarded items in chosen minus unchosen pairs). We found that the two difference scores were significantly correlated (Bayesian linear regression predicting choice difference scores as a function of memory difference scores: $\beta_{\text{memory}} = -0.38$ [$-0.60, -0.17$]), suggesting that the more memory was impaired, the less people tended to inversely infer the value of unchosen items (Figure 2D). We ran this same analysis in the data from the Pilot study ($n = 108$), where the effect did not reach significance ($\beta_{\text{memory}} = -0.37$ [$-0.80, 0.06$]), perhaps due to lack of power.

A Policy Gradient Model Captures the Memory-Based Inverse Decision Bias

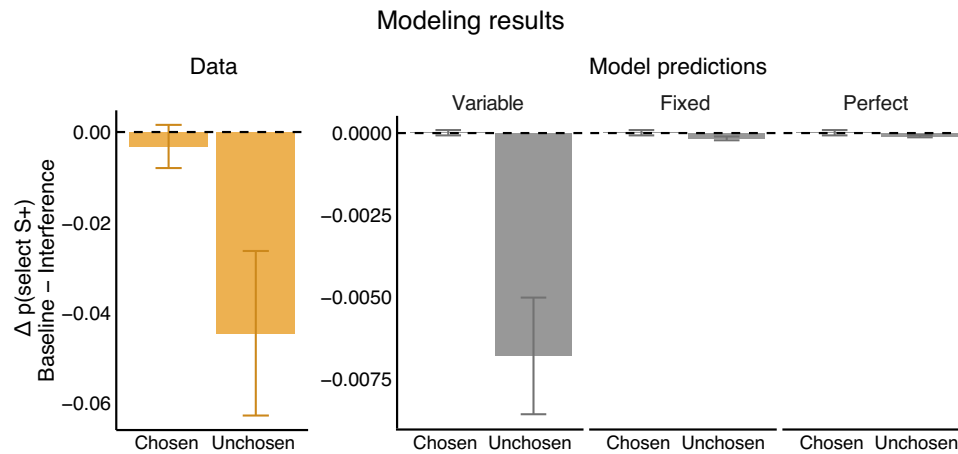
Figure 3 shows the behavior of the three models described in the Methods, after fitting them to the data. The key finding in Experiment 1 is a significant difference in $p(\text{select } S+)$ for Interference vs. Baseline conditions, in unchosen pairs, but not

Figure 2
Associative Memory Modulates the Inverse Inference of Value



Note. (A) The associative memory manipulation was successful at modifying the memory of choice options. Presenting the chosen items of the one half of Deliberation pairs with novel items (Interference condition, orange) was intended to induce retroactive interference, thereby decreasing the association between choice options. Repeating the other half of the Deliberation pairs (Baseline condition, turquoise) was intended to serve as a baseline condition. The manipulation resulted in worse memory accuracy for Interference pairs compared to Baseline pairs, tested in the Surprise Memory phase. (B) The memory manipulation affected the inverse inference of value of unchosen options. When faced with a choice between two previously chosen stimuli, for which value was explicitly shown [S_{chosen} (learned)], participants tended to select the rewarded option ($S_{\text{chosen}}+$) in both Baseline and Interference conditions, suggesting they successfully learned their values and the manipulation did not modulate direct learning of outcomes. For pairs of previously unchosen options, which were never directly associated with any reward [S_{unchosen} (inferred)], participants tended to select the option previously associated with an unrewarded item (S_{unchosen}^0) in Baseline pairs, but not in Interference pairs, suggesting the manipulation modulated only inferred outcomes. (C) Associative memory was related to inverse decision bias (difference in mean probability to choose rewarded items for chosen and unchosen pairs) regardless of experimental condition. (D) The effect of interference on associative memory (difference between turquoise and orange bars in panel A, higher scores depict stronger interference effect) was related to its effect on counterfactual inference (difference between turquoise and orange bars in panel B for unchosen pairs, lower scores depict stronger interference effect). In all panels, points denote trial-averaged data of individual participants. In panels A and B, error bars denote *SEM*. In panel C, the lines denote the fit of a Bayesian linear regression predicting inverse decision bias as a function of memory accuracy and condition. Model coefficients were rearranged to produce separate regression lines for every experimental condition, each predicting inverse bias as a function of memory. Model fits and memory coefficients (beta) depict median and 95% highest density interval estimates. If viewing in grayscale, orange equals light gray, and turquoise equals dark gray. See the online article for the color version of this figure.

Figure 3
The Variable-Memory Model Best Predicts the Key Empirical Finding in Experiment 1



Note. Gray bars depict model predictions of the three memory models. Each bar shows the model-based interference effect on choice behavior. That is, the mean difference between the probability to select a rewarded item in Baseline minus Interference conditions (lower scores depict a stronger interference effect), in chosen and unchosen pairs. Yellow bars depict the corresponding behavior using choices from the Final Decisions phase in Experiment 1. Error bars show *SE*. If viewing in grayscale, gray equals dark gray, and yellow equals light gray. See the online article for the color version of this figure.

in chosen ones. All three models capture this behavior for chosen pairs, showing no difference between the two conditions in the tendency to select rewarded items (all p s $> .8$). Importantly, for unchosen pairs, the variable memory model mirrors the experimental data, showing a significant positive difference value between conditions, $t(314) = -3.82$, $p = .0002$, Cohen's $d = 0.22$. The fixed memory and perfect memory models showed a smaller effect (fixed: $t(314) = -2.79$, $p = 0.006$, Cohen's $d = 0.16$, perfect: $t(314) = -2.63$, $p = 0.009$, Cohen's $d = 0.15$).

To compare the three models and evaluate the degree to which they agree with the data, we applied random-effects Bayesian model comparison (Rigoux et al., 2014), which allows us to estimate the frequency with which each model occurs within the population. Applying this method to the Final Decisions' choices (which were not used to fit the models), we found that the expected frequency of the variable memory model was 0.92, with a protected exceedance probability of greater than .999. Thus, the variable model does a much better job than the other two models at accounting for the key behavioral pattern—that is, the difference between the Baseline and Interference conditions (Figure 3). At the same time, it may be worth noting that this model does not fully capture the quantitative effect size observed in this dataset.

In summary, we can successfully formalize the memory-dependence of policy updating using a policy gradient model with memory that varies across items. Importantly, this variability allows us to account for the effectiveness of our interference manipulation in ways that the pre-existing policy gradient model (the “perfect” model) does not.

Discussion

This study provides empirical evidence for the role of memory in counterfactual value assignment as well as a computational framework

for considering this behavioral pattern. The study replicates and extends our previous findings showing that people tend to inverse the value of unchosen options relative to the outcome of a chosen option, a phenomenon we refer to as inverse decision bias (Biderman & Shohamy, 2021). Here we extend this result by further probing the role of memory in this phenomenon, showing that a manipulation of memory—before value learning even takes place—has a subsequent impact on the inverse decision bias. Finally, we tie the empirical findings to a new theoretical account, showing that the behavioral patterns can be understood in the context of a memory-based policy gradient model. Together, the behavioral pattern and the model highlight the significant role of associative memory in the inverse decision bias, pointing to a connection between memory, counterfactual reasoning, and theoretical models of learning and decision-making.

Our findings have significant implications for models of learning and decision-making. Participants in the study receive explicit reinforcement regarding the value of the chosen options and are later asked to make decisions about the options they did not choose and for which value was not shown. As such, behavior on this task provides an example of value learning that cannot be derived solely through direct reinforcement, as predicted by value-based models like Q-learning (Watkins & Dayan, 1992; Sutton & Barto, 1998; Bennett et al., 2021). Policy gradient models, on the other hand, are not driven by direct learning of value, but by the updating of model parameters according to a behavioral policy. For binary decisions and a logistic policy, after the outcomes of the chosen option are revealed, the value of both the chosen and unchosen options is updated in an inverse manner, mirroring the empirical pattern of behavior displayed on this task.

The standard policy gradient model, however, fails to capture a critical and robust pattern in our data—the relationship between the inverse decision bias and associative memory of choice options. To adjust the value of the unchosen option in relation to the chosen option, one has to remember that the two options were once

considered together. Indeed, Wang and colleagues have included memory components in policy gradient models (an experience replay buffer, storing memories that can later be entered into the policy gradient update), and showed improved performance in different simulated environments (Wang et al., 2016). Here, we generalized the standard logistic policy gradient model to include a memory parameter that controls the magnitude of the inverse value update. Our main goal with the modeling here was to provide a conceptual framework for understanding the empirical observations, rather than providing a quantitative assessment *per se*. Nonetheless, we compared three models with different memory components and found that a model that allows memory to vary within participants captures the key finding of our experiment, namely, that the inverse update of unchosen options is reduced when memory is interfered.

The past several years have seen major progress in understanding the role that memory plays in decision-making (for reviews see Biderman et al., 2020; Palombo et al., 2015). The current study extends previous work by providing evidence for a significant role of memory in counterfactual reasoning, suggesting that memory modulates the manner in which people update value for unrealized options. Importantly, the connection between memory and inverse decision bias provides a framework to consider counterfactual reasoning not as binary—you either think about unchosen options or not—but as vulnerable to factors that affect our memory. While the current study focused on retroactive interference, which reduced memory and altered the magnitude of the inverse decision bias, we assume that any number of factors known to affect memory will also, similarly, affect the inverse decision bias. For example, memory is known to be affected by novelty (Kafkas & Montaldi, 2018; Tulving et al., 1996), temporal order (Capitani et al., 1992; Howard & Kahana, 2002; Murdock, 1962), spacing in learning (Cepeda et al., 2006; Ebbinghaus, 2013), activeness of learning (Markant et al., 2016), just to name a few. Future research will be necessary to determine which of these factors may also affect the way people make predictions about unchosen options.

The current study targeted associative memory of choice options because we reasoned that to infer the value of the unchosen option in relation to the chosen option, one had to form an association between them. To weaken the association between choice options we induced retroactive interference by coupling one of the options with a new option. Manipulated choice options resulted in reduced memory and reduced inverse decision bias for later choices. While retroactive interference is a standard method used to reduce associative memory (Jenkins & Dallenbach, 1924; McGeoch, 1932; Sosic-Vasic et al., 2018), an effect which we replicate here, the direction of the manipulation in our own data could also run the other way. That is, given the literature showing that choices can alter memory (DuBrow et al., 2019; Murty et al., 2015, 2019), it is possible that the retroactive interference manipulation reduced the inverse decision bias and that, in turn, changed associative memory. Either way, our findings show evidence for a tight relationship between memory and counterfactual decisions and that using a manipulation chosen for its previously established effects on memory results in changes to both measurements.

Indeed, we found that our manipulation modulated decisions in trials of unchosen items, for which value could only be inferred, but it did not affect decisions in trials of chosen items, for which value was directly learned. These findings echo previous studies showing that memory mechanisms are involved in inference and

generalization (Heckers et al., 2004; Preston et al., 2004; Schuck & Niv, 2019; Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012; Wong et al., 2019; Zeithamova et al., 2012), and raise new questions about how we infer value for choice options which were never directly experienced.

Altogether, this study links two separate literatures—counterfactual reasoning and models of learning and decision-making—both of which predict an inverse inference for unchosen options but arriving at this prediction from different perspectives. From the point of view of studies of counterfactual reasoning, the inverse nature of value inference has been tied to the behavioral tendency to continue to separate the value of options even after the decision was made. Studies of choice-induced preference change (e.g., Bem, 1967; Festinger, 1957; Sharot et al., 2010; Voigt et al., 2017) show that for two equally valued options, the mere act of deciding—even without feedback—induces a tendency to overvalue the chosen option and undervalue the unchosen one. Similarly, studies of choice-supportive memory (Henkel & Mather, 2007; Lind et al., 2017; Mather et al., 2000) show that participants tend to better learn and remember the positive aspects of their chosen option and the negative aspects of their unchosen one, thereby increasing the contrast of value between the options. Importantly, our study diverges from choice-supportive memory and choice-induced preference change in two meaningful ways. First, our study focuses on how value is changing as a function of learning from subsequent outcomes, rather than as a function of the decision alone. Furthermore, our account is bidirectional: the chosen option is not always preferred. Depending on the outcomes associated with the chosen option, the unchosen option may be inferred to be better.

Policy gradient models offer a different perspective on why unchosen options are inversely inferred. The standard logistic policy model is parameterized by the difference between the two options' action propensities. Taking a gradient over the policy parameters results in an inverse value update of the two options. In other words, the comparison between the two options made during the decision process does not only regulate the decision itself but also affects the shape of the gradient over the policy parameters, and consequently the value update of the action propensities. The connection between the psychological and computational theories provides a quantitative framework for future experiments and theories on the consequences of contrasting two options to each other.

Context

Our decisions reflect who we are and shape who we become. There has been enormous progress in understanding the role of learning and reinforcement in decision-making, stimulated by converging evidence from computational models, behavior, and neural measurements. Together, this work has uncovered how outcomes of past decisions are used to update value representations, thereby guiding future decisions. However, many decisions cannot be based on value that was learned from direct reinforcement. For example, we often make choices among options we never experienced before. These decisions rely on generalization from memory or inferential reasoning, yet relatively little is known about the mechanisms involved. The current study adds to a growing literature exploring the deep connections between associative and episodic memory and decision-making. Here, we focus on the role of memory in counterfactual reasoning in support of decisions about previously

unchosen options. We show that a particular class of models—policy gradient reinforcement learning models—account for a behavioral inverse decision bias and that the bias and the policy itself can be understood as emerging from memories formed during the process of deliberation. The hypothesis, design, and the model reflect the combined influence of two subfields of psychological science: memory and decision-making—showcasing the value of considering their combined influence.

References

- Balleine, B. W. (2018). The motivation of action and the origins of reward. In R. Morris, A. Borstein, & A. Shenhav (Eds.), *Goal-directed decision making: Computations and neural circuits* (pp. 429–455). Elsevier. <https://doi.org/10.1016/B978-0-12-812098-9.00019-X>
- Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200. <https://doi.org/10.1037/h0024835>
- Bennett, D., Niv, Y., & Langdon, A. J. (2021). Value-free reinforcement learning: Policy optimization as a minimal model of operant behavior. *Current Opinion in Behavioral Sciences*, *41*, 114–121. <https://doi.org/10.1016/j.cobeha.2021.04.020>
- Bideman, N., Bakkour, A., & Shohamy, D. (2020). What are memories for? The hippocampus bridges past experience with future decisions. *Trends in Cognitive Sciences*, *24*(7), 542–556. <https://doi.org/10.1016/j.tics.2020.04.004>
- Bideman, N., Gershman, S. J., & Shohamy, D. (2022). *The role of memory in the inverse inference of value of unchosen options*. osf.io/qad57
- Bideman, N., & Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, *12*(1), Article 4648. <https://doi.org/10.1038/s41467-021-24907-x>
- Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software* *80*(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Capitani, E., Della Sala, S., Logie, R. H., & Spinnler, H. (1992). Recency, primacy, and memory: Reappraising and standardising the serial position curve. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *28*(3), 315–342. [https://doi.org/10.1016/S0010-9452\(13\)80143-8](https://doi.org/10.1016/S0010-9452(13)80143-8)
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. <https://doi.org/10.1037/0033-2909.132.3.354>
- Champely, S. (2020). *pwr: Basic functions for power analysis*. R package version 1.3.0. <https://CRAN.R-project.org/package=pwr>
- Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality & Social Psychology Bulletin*, *25*(2), 215–224. <https://doi.org/10.1177/0146167299025002007>
- Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems* (p. 460). Massachusetts Institute of Technology Press.
- de Leeuw, J. R. (2015). Jspych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- DuBrow, S., Eberts, E. A., & Murty, V. P. (2019). A common mechanism underlying choice's influence on preference and memory. *Psychonomic Bulletin & Review*, *26*(6), 1958–1966. <https://doi.org/10.3758/s13423-019-01650-5>
- Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, *13*(6), 513–519. <https://doi.org/10.1111/1467-9280.00490>
- Durkin, C., Hartnett, E., Shohamy, D., & Kandel, E. R. (2020). An objective evaluation of the beholder's response to abstract and figurative art based on construal level theory. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(33), 19809–19815. <https://doi.org/10.1073/PNAS.2001772117>
- Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences*, *20*(4), 155–156. <https://doi.org/10.5214/ans.0972.7531.200408>
- Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.
- Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). *rstanarm: Bayesian applied regression modeling via stan*. R package version 2.21.1. <https://mc-stan.org/rstanarm>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis: Third edition* (2nd ed.). Academic Press. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36. <https://doi.org/10.1006/anbe.2003.2174>
- Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476. <https://doi.org/10.1038/nature20101>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, *14*(2), 153–162. <https://doi.org/10.1002/hipo.10189>
- Henkel, L. A., & Mather, M. (2007). Memory attributions for choices: How beliefs shape our memories. *Journal of Memory and Language*, *57*(2), 163–176. <https://doi.org/10.1016/j.jml.2006.08.012>
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. <https://doi.org/10.1006/JMPS.2001.1388>
- Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *35*(4), 605–612. <https://doi.org/10.2307/1414040>
- Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4–6), 535–547. [https://doi.org/10.1016/S0893-6080\(02\)00047-3](https://doi.org/10.1016/S0893-6080(02)00047-3)
- Kafkas, A., & Montaldi, D. (2018). How do memory systems detect and respond to novelty? *Neuroscience Letters*, *680*, 60–68. <https://doi.org/10.1016/j.neulet.2018.01.053>
- Kahnt, T., Park, S. Q., Burke, C. J., & Tobler, P. N. (2012). How glitter relates to gold: Similarity-dependent reward prediction errors in the human striatum. *Journal of Neuroscience*, *32*(46), 16521–16529. <https://doi.org/10.1523/JNEUROSCI.2383-12.2012>
- Kahnt, T., & Tobler, P. N. (2016). Dopamine regulates stimulus generalization in the human hippocampus. *eLife*, *5*, Article e12678. <https://doi.org/10.7554/eLife.12678>
- Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience*, *31*(14), 5504–5511. <https://doi.org/10.1523/JNEUROSCI.6316-10.2011>
- Liashenko, A., Dizaji, A. S., Melloni, L., & Schwiedrzik, C. M. (2020). Memory guidance of value-based decision making at an abstract level of representation. *Scientific Reports*, *10*(1), Article 21496. <https://doi.org/10.1038/s41598-020-78460-6>

- Lind, M., Visentini, M., Mäntylä, T., & Del Missier, F. (2017). Choice-supportive misremembering: A new taxonomy and review. *Frontiers in Psychology*, 8, Article 2062. <https://doi.org/10.3389/fpsyg.2017.02062>
- Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6, Article 1171. <https://doi.org/10.3389/fpsyg.2015.01171>
- Macmillan, N. A., & Douglas Creelman, C. (1991). *Detection theory: A user's guide*. CUP Archive. <https://play.google.com/store/books/details?id=Pfw3AAAAIAAJ>
- Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *The Journal of Open Source Software*, 3(22), 470–471. <https://doi.org/10.21105/joss.00470>
- Marciano-Romm, D., Romm, A., Bourgeois-Gironde, S., & Deouell, L. Y. (2016). The alternative omen effect: Illusory negative correlation between the outcomes of choice options. *Cognition*, 146, 324–338. <https://doi.org/10.1016/j.cognition.2015.10.003>
- Markandya, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, 10(3), 142–152. <https://doi.org/10.1111/mbe.12117>
- Mather, M., Shafir, E., & Johnson, M. K. (2000). Misrememberance of options past: Source monitoring and choice. *Psychological Science*, 11(2), 132–138. <https://doi.org/10.1111/1467-9280.00228>
- McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, 39(4), 352–370. <https://doi.org/10.1037/h0069819>
- McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in science conference* (Vol. 445, pp. 51–56).
- Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482–488. <https://doi.org/10.1037/h0045106>
- Murty, V. P., DuBrow, S., & Davachi, L. (2015). The simple act of choosing influences declarative memory. *Journal of Neuroscience*, 35(16), 6255–6264. <https://doi.org/10.1523/JNEUROSCI.4181-14.2015>
- Murty, V. P., DuBrow, S., & Davachi, L. (2019). Decision-making increases episodic memory via postencoding consolidation. *Journal of Cognitive Neuroscience*, 31(9), 1308–1317. https://doi.org/10.1162/jocn_a_01321
- Navarro, D. (2013). *Learning statistics with R: A tutorial for psychology students and other beginners: Version 0.5*. University of Adelaide Adelaide.
- O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, 14(6), 769–776. <https://doi.org/10.1016/j.conb.2004.10.016>
- O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, 68(1), 73–100. <https://doi.org/10.1146/annurev-psych-010416-044216>
- Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6(1), Article 8096. <https://doi.org/10.1038/ncomms9096>
- Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, 13(8), Article e1005684. <https://doi.org/10.1371/journal.pcbi.1005684>
- Palombo, D. J., Keane, M. M., & Verfaellie, M. (2015). How does the hippocampus shape decisions? *Neurobiology of Learning and Memory*, 125, 93–97. <https://doi.org/10.1016/j.nlm.2015.08.005>
- Peters, J., & Schaal, S. (2006). Policy gradient methods for robotics. *IEEE International Conference on Intelligent Robots and Systems* (pp. 2219–2225). <https://doi.org/10.1109/IROS.2006.282564>
- Pompilio, L., & Kacelnik, A. (2005). State-dependent learning and suboptimal choice: When starlings prefer long over short delays to food. *Animal Behaviour*, 70(3), 571–578. <https://doi.org/10.1016/j.anbehav.2004.12.009>
- Pompilio, L., Kacelnik, A., & Behmer, S. T. (2006). State-dependent learned valuation drives choice in an invertebrate. *Science*, 311(5767), 1613–1615. <https://doi.org/10.1126/science.1123924>
- Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. E. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus*, 14(2), 148–152. <https://doi.org/10.1002/hipo.20009>
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, 84, 971–985. <https://doi.org/10.1016/j.neuroimage.2013.08.065>
- Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, 364(6447), Article eaaw5181. <https://doi.org/10.1126/science.aaw5181>
- Sharot, T., Velasquez, C. M., & Dolan, R. J. (2010). Do decisions shape preference? Evidence from blind choice. *Psychological Science*, 21(9), 1231–1235. <https://doi.org/10.1177/0956797610379235>
- Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron*, 60(2), 378–389. <https://doi.org/10.1016/j.neuron.2008.09.023>
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569. <https://doi.org/10.1177/0956797614567341>
- Sosic-Vasic, Z., Hille, K., Kröner, J., Spitzer, M., & Kornmeier, J. (2018). When learning disturbs memory—temporal profile of retroactive interference of learning on memory formation. *Frontiers in Psychology*, 9, 82. <https://doi.org/10.3389/fpsyg.2018.00082>
- Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT Press Cambridge.
- Tulving, E., Markowitsch, H. J., Craik, F. I. M., Habib, R., & Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, 6(1), 71–79. <https://doi.org/10.1093/cercor/6.1.71>
- Van Rossum, G., & Drake, F. L. (2000). *Python reference manual*. iUniverse Indiana.
- Voigt, K., Murawski, C., & Bode, S. (2017). Endogenous formation of preferences: Choices systematically change willingness-to-pay for goods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(12), 1872–1882. <https://doi.org/10.1037/xlm0000415>
- Wang, Z., Bapst, V., Heess, N., Mnih, V., & Munos, R. (2016). Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*.
- Watkins, C. C. H., & Dayan, P. (1992). Q-Learning. *Machine Learning*, 8(3–4), 279–292. <https://doi.org/10.1007/BF00992698>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
- Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations*. R package version 1.4.0. <https://CRAN.R-project.org/package=stringr>.
- Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation*. R package version 1.0.7. <https://dplyr.tidyverse.org>
- Wickham, H., Vaughan, D., & Girlich, M. (2023). *tidyr: Tidy messy data*. <https://github.com/tidyverse/tidyr>
- Wilke, C. (2020). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. R package version 1.1.1. <https://CRAN.R-project.org/package=cowplot>
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4), 229–256. <https://doi.org/10.1007/BF00992696>
- Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, 338(6104), 270–273. <https://doi.org/10.1126/science.1223252>

- Wong, F. S., Westbrook, R. F., & Holmes, N. M. (2019). “Online” integration of sensory and fear memories in the rat medial temporal lobe. *eLife*, 8, Article e47085. <https://doi.org/10.7554/eLife.47085>
- Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, 75(1), 168–179. <https://doi.org/10.1016/j.neuron.2012.05.010>
- Zhu, H. (2021). *kableExtra: Construct complex table with ‘kable’ and pipe syntax*. R package version 1.3.4. <https://CRAN.R-project.org/package=kableExtra>

Received March 14, 2022

Revision received December 2, 2022

Accepted December 18, 2022 ■