bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

What's past is present: Reminders of past choices bias decisions for reward in humans.

Aaron M. Bornstein [1,*], Mel W. Khaw [3], Daphna Shohamy [4] & Nathaniel D. Daw [1,2]

1. Neuroscience Institute and 2. Department of Psychology, Princeton University, Princeton, NJ, USA.

3. Center for Neural Science, New York University, New York, NY, USA.

4. Department of Psychology; Zuckerman Mind, Brain, Behavior Institute;

and Kavli Center for Brain Science, Columbia University, New York, NY, USA.

* To whom correspondence should be addressed: aaronmb@princeton.edu

Abstract

We provide evidence that decisions are made by consulting memories for individual past experiences, and that this process can be biased in favor of past choices using incidental reminders. First, in a standard rewarded choice task, we show that a model that estimates value at decision-time using individual samples of past outcomes fits choices and decision-related neural activity better than a canonical incremental learning model. In a second experiment, we bias this sampling process by incidentally reminding participants of individual past decisions. The next decision after a reminder shows a strong influence of the action taken and value received on the reminded trial. These results provide new empirical support for a decision architecture that relies on samples of individual past choice episodes rather than incrementally averaged rewards in evaluating options, and has suggestive implications for the underlying cognitive and neural mechanisms.

How does experience guide our choices? One approach is to use a simple heuristic: do what has worked best recently. This idea is captured in prominent reinforcement learning (RL) models, which build a running average of the rewards received for each action taken, updating this average incrementally as new rewards are experienced. Such averages can be updated by error-driven learning, tying this approach to observations of reward prediction errors in the brain [1, 2]. This approach is effective and efficient in the sorts of tasks typically used to study reward learning, which usually involve many repeated choices among a small number of options whose only salient difference is how much or how often they are rewarded [2–5]. A running average falls short, however, in more naturalistic settings, such as those that involve generalizing from past experience to make decisions with little or no direct experience of reward to guide them [6–13].

Here we provide evidence in favor of a more flexible approach in which choosers draw on memories for individual instances of relevant previous choices and use them to predict how the current decision might turn out. For example, when deciding where to eat dinner, I might recall a particular evening dining at a particular restaurant, and make my decision based on how much I enjoyed the meal I had that night. Making decisions this way allows us to pick and choose the most relevant of previous experiences, rather than relying on static summary representations. The latter would be useless in considering a brand new restaurant, whereas a richer memory representation of each past experience could be used to develop an informed guess, by generalizing from experiences with similar establishments. In this way, evaluating options by sampling resembles nonparametric and kernel-based methods long employed in statistics and machine learning, which draw on individual samples of raw experience [14, 15]. The effectiveness of these estimation algorithms reinforces recent interest in the idea that humans and animals draw on memory to flexibly compute decision variables at choice time [12, 16–21], rather than relying on precomputed averages.

We present two empirical results in support of this hypothesis. First, we show that a sampling model is a better fit to trial-by-trial choice sequences than incremental learning, even in the sort of repeated choice task

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

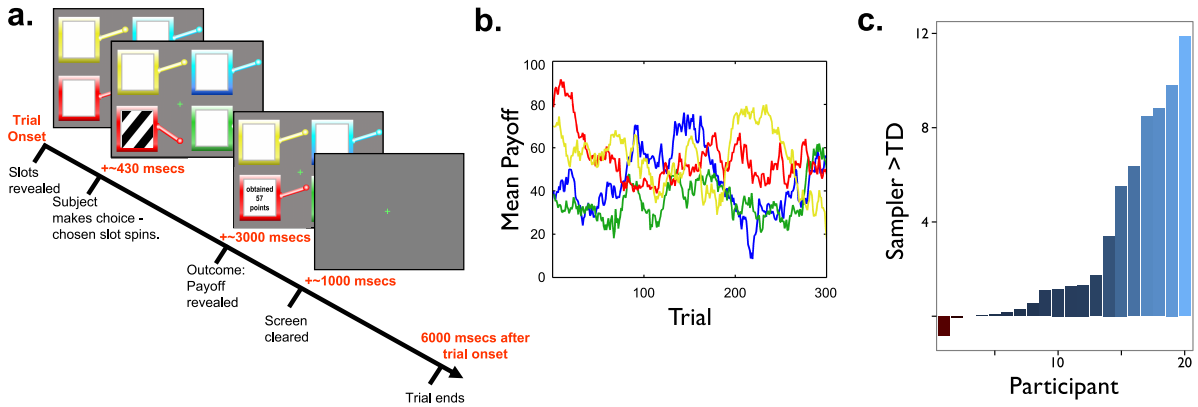

Figure 1: **Restless bandit task** (from Daw et al., 2006) a. **Four-armed bandit**. Participants chose between four slot machines to achieve points. b. **Payoffs**. The mean amount of points paid out by each machine varied slowly over the course of the experiment. c. **Model comparison**. Log Bayes Factors favoring sampling over the TD model.

previously studied using RL models. In a second study, we directly test the prediction that choices draw on memories of individual trials, by demonstrating a causal effect on choices of specific, selected experiences, brought to mind by refreshing memories of individual past decision trials.

Results

We first formally investigated whether a sampling model could provide a better trial-by-trial fit to human choices than standard incremental learning. To test this, we re-analyzed choice and neuroimaging data from a previously published learning study [4]. Twenty participants (14 presented with neuroimaging data previously, six additional behavior-only participants included here) performed a series of choices between four virtual slot machines with time-varying payoffs (Figure 1a,b). Here, we compared a standard incremental learning rule (of the form tested in the original study) to a sampling model that evaluates each option by retrieving the rewards received at individual past choices of that option, stochastically sampled with probabilities given by their temporal recency (see *Methods*). These sampled rewards are averaged to compute a net value for each option. In the limit, when a potentially infinite number of samples may be considered, these sample-based averages are identical to those computed by a standard running average model. When fewer samples are considered, the models make similar predictions on average, but the range of resulting choices differs systematically. For instance, the running average approach predicts a fixed, vanishingly small influence of long-past rewards, while in the sampling model events from the far past could still have occasional, but sizable, influence on choices. Our primary model of interest was a sampling model that drew only one sample of past experience. (We also considered variants that drew a greater number of samples, none of these were superior to the one-sample version.) Critically, the two models were matched in all other respects; in addition to recency weighting, the choice rules and free parameters were identical.

We computed, for each trial and each participant individually, the probability that each model would produce the choice observed, and estimated for each model the parameter values that maximized these choice likelihoods. Comparing these maximum-likelihood estimates we found that, across the population and for eighteen of twenty participants individually, choices were better fit by our sampling model than by the learning model (mean log Bayes Factor 3.05; Figure 1c).

Given the Sampler's superior fit to behavior, we next asked whether the expectations and learning signals observed in neural activity during decisions were better modeled using the Sampler model. Specifically, we tested whether the well-studied neural correlates of key decision variables – chosen value and Reward

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

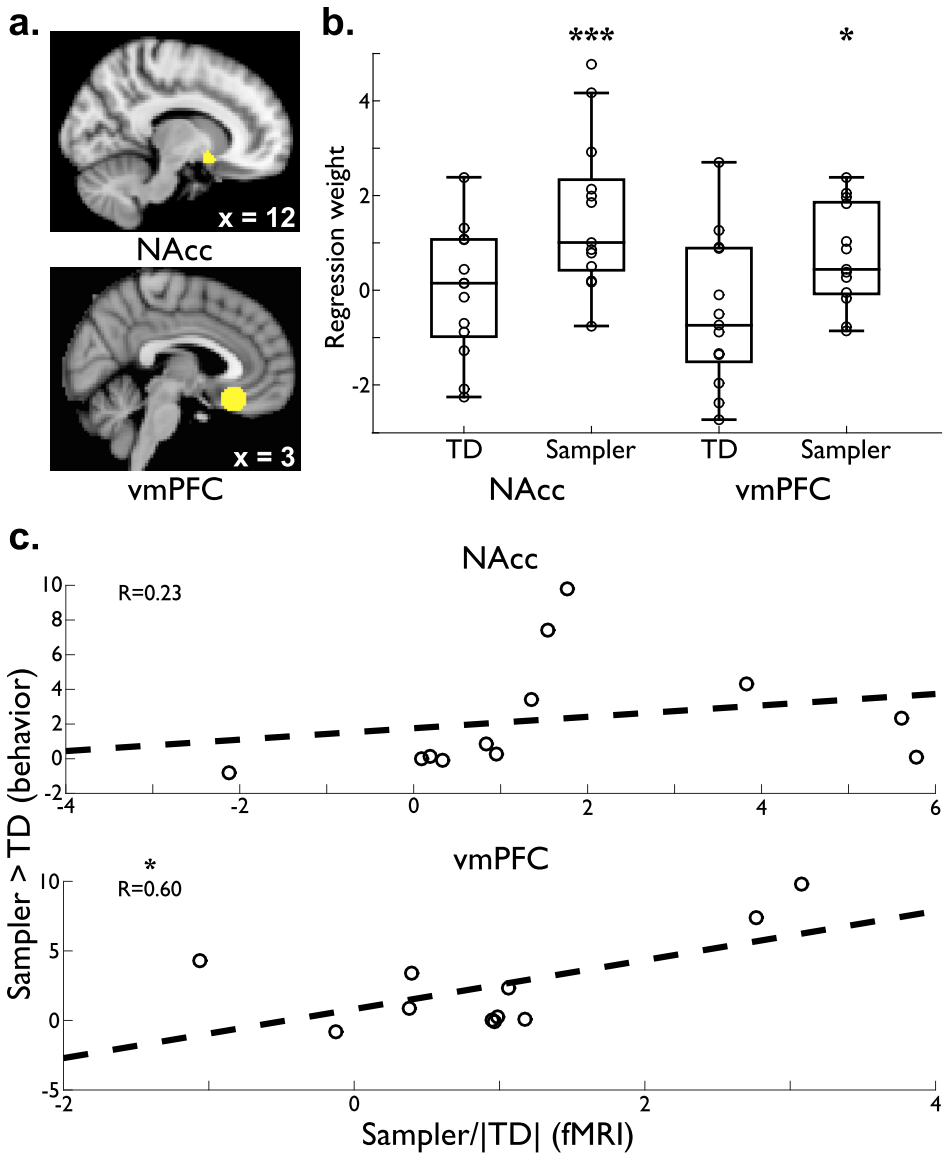

Figure 2: **Sampling fit to neural decision variables** a. **Regions of interest**. We isolated voxels of interest that corresponded to previous reports of the neural substrates for the decision variables analyzed here: Chosen Value (CV), in ventromedial prefrontal cortex (vmPFC), and Reward Prediction Error (RPE), in the nucleus accumbens (NAcc). b. **Simultaneous regression**. Candidate timeseries for each decision variable were generated according to each of the two models, and entered into a simultaneous regression against the BOLD timeseries extracted from the relevant ROI. Each plotted point represents the regression coefficient for the respective model-timeseries pair; box plots display the mean and interquartile range (* $p < 0.05$, *** $p < 0.005$). Both regressions support the hypothesis that the Sampler model underlies neural signals (NACC-RPE: $t_{13} = 3.5409, p = 0.004$; vmPFC-CV: $t_{13} = 2.4112, p = 0.033$). c. **Neural fit correlated with behavioral fit**. The relative regression coefficient relating the Sampler model to the BOLD timeseries were correlated with the log Bayes Factor describing Sampler's fit to behavior. We used robust correlations to infer trends irrespective of outliers; the outliers excised in each correlation (one for NAcc, two for vmPFC) are left unplotted here. In both regions, the correlation between the behavioral and fMRI effects was strong and consistent with a greater tendency for Sampler to better explain neural signal in cases where it better explained behavior (NAcc: Robust Pearson correlation $R = 0.2274$, $p = 0.2836$; vmPFC: $R = 0.5979$, $p = 0.0282$).

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Prediction Error (RPE) – were better predicted by the sampling model than by TD. We first identified regions of interest (ROIs) encompassing areas previously shown to reflect this activity: ventromedial prefrontal cortex / medial orbitofrontal cortex (hereafter: vmPFC) for chosen value, and Nucleus Accumbens (NAcc) for RPE. Extracting the timeseries of activity within these ROIs, we next performed a simultaneous regression containing the timeseries of variables predicted by both models, along with several regressors of no interest. Comparing the distribution of resulting per-participant regression weights against zero using a two-tailed, one-sample t-test, we evaluated whether each model was a significant predictor of the target BOLD timeseries. In both cases, the predictions of the sampling model captured additional variance in the BOLD timeseries that was not modeled by TD (Figure 2b). Across participants, regression on the NAcc timeseries revealed significant contribution of the RPE variable, as generated by the Sampler model ($t_{13} = 3.5409, p = 0.004$). Similarly, the regression weights on the Sampler model's CV regressor showed a positive trend towards explaining significant variance in the vmPFC timeseries ($t_{13} = 1.8352, p = 0.0914$). Importantly, because of the way multiple regression competes between predictor variables to explain variance, these results imply that decision variables defined by the Sampling model significantly explained additional variance over and above any explained by the TD model, but the converse was not true.

Lastly, we investigated whether the Sampler model's fit to a participant's neural timeseries corresponded to how well it captured their behavior. We performed a robust correlation analysis [22] comparing the ratio of the regression weights computed for the models on neural activity (Sampler/—TD—) to the log Bayes Factor in favor of the Sampler model on choices, as estimated above. We found that, on balance, participants whose neural timeseries were better explained by the Sampler model also had choices better captured by the Sampler model (Figure 2c). In the NAcc, the robust correlation analysis resulted in a robust Pearson correlation coefficient of $R = 0.2274$ ($t_{11} = 0.7744, p = 0.2836$); in vmPFC the correlation reached statistical significance ($R = 0.5979, t_{10} = 2.4738, p = 0.0282$).

These results concords with our hypothesis that human decisions are guided by samples of individual past choices, even in a simple repeated choice tasks. However, one disadvantage of applying sampling models to this type of task is that trials are essentially similar; we cannot directly observe which individual trials participants have sampled. Moreover, the experiment was correlational, rather than causal. To address this, we performed a second behavioral experiment to provide more direct support for the hypothesis by bringing the sampling process under experimental control, so as to measure the impact of a single, selected experience on choices. We leveraged the fact that multiple aspects of the choice experience may be bound together as a single representation. Therefore we used choice-incidental – but still trial-unique – information to tag each choice as a unique event. Specifically, we modified the bandit task from Experiment 1 so that each trial involved a uniquely identifying photograph of an everyday object – a "ticket" emitted by the chosen slot machine (see also [23]). To simplify the task and analysis, the number of choice options was reduced to two from four, and outcome values were limited to wins or losses of $5 (Figure 3a). The probability that each machine would pay a winning ticket varied from trial to trial (Figure 3b). Other than the ticket presentation, this task matches the sort of "two-armed bandit" traditionally employed in reinforcement learning studies.

Interspersed among the 130 choices were 32 recognition probe trials on which participants were asked whether or not they recognized a given "ticket" image (Figure 3c). These probes were intended to bring to mind the specific trial on which the probed image was first experienced – including the bandit chosen before the ticket appeared and the outcome received after. We hypothesized that this reminder would make the original experience more likely to be sampled during the ensuing choice. In contrast, standard running average models predict no such effect, since they maintain and base choices only on the summary statistics. Probes evoked choices that were, on average, 39 trials in the past – a temporal horizon minimizing both the influence of that reward on a recency weighted running average, as well as the likelihood that the reminded trial would still be present in working memory [24]. After each probe the choice task continued as before.

Matching the first analysis, the sampling model proved a superior fit to choices than did an incremental learning model for 20 out of 21 participants individually and across the population as a whole (mean log Bayes Factor 3.46; Figure 4a). However the key test of our hypothesis is whether choices following a recognition probe show an effect of the cued trial. For example, if a given recognition memory probe evoked a trial on which the participant chose the blue bandit and was rewarded, then the participant should be more likely to

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

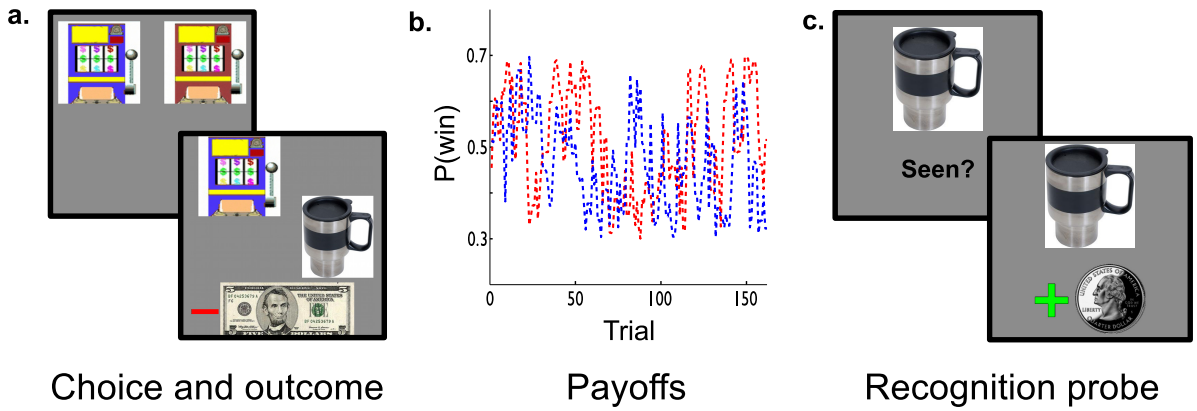

Figure 3: **Ticket bandit task**. a. **The ticket-bandit task**. Each slot machine ("bandit") delivered tickets – trial-unique photographs – associated with a dollar value – either -$5 or $5. b. **Payoff probabilities**. The probability of each bandit paying out a winning ticket varied slowly over the course of the experiment. Participants were told that their total payout would be contingent both on the number of winning tickets they accrued and also their ability to correctly respond on a post-task memory test asking them to recall the reward value and slot machine associated with each ticket. c. **Memory probes**. Participants encountered 32 recognition memory probes. On 26 of these probe trials, participants were shown objects that were either received on a previous choice trial ("valid"), while on others they were shown new objects that were not part of any previous trial ("invalid"). Participants were asked only to perform a simple old/new recognition judgement – to press "yes" if they had seen the image previously in this task, and "no" if they had not. After each recognition probe, the sequence of slot machine choices continued as before.

choose the blue bandit on the subsequent choice trial. Conversely, if the choice had resulted in a loss, then they should be more likely to choose the alternative bandit (in this example, red). Incremental RL would be unaffected by the memory probes because it does not maintain memories for individual trials and because the memory probe trials do not provide relevant direct reward experience that could be used to update running average values for the slot machines. Consistent with the sampling model, we observed that choices following a memory probe were also significantly influenced by the much older experience evoked by the probed ticket ($p < 0.004$; Figure 4b). The magnitude of the increase in choice probability was comparable to that of a reward directly received just two trials in the past, suggesting that the reminder probe was effective in bringing to mind the associated bandit-reward link, and that reward information then incorporated into choice, on a large fraction of post-reminder trials.

Discussion

These results establish that choices can be made by sampling from individual trial memories, an architecture wholly different from those traditionally employed to study reward learning. We showed that this model can capture the characteristic – and behaviorally observed [25] – recency dependence of incremental averaging approaches, yet provides a superior fit to behavior and decision-related neural signals even in a task originally designed to induce such incremental value updating. This finding goes beyond earlier work that showed that aggregate features of choices – such as a chooser's variable sensitivity to risky options – are also consistent with a similar sampling model [26, 27]. However none of these results directly expose the key hypothesis that choices are driven by memories of individual trials. We therefore tested this idea more directly in a second experiment. We showed that choices can be influenced by bringing to mind individual past choice experiences using incidental stimulus-stimulus associations present at the time of the original decision. Episodic memory is characterized by the formation of stimulus-stimulus associations during a single experience. Therefore

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

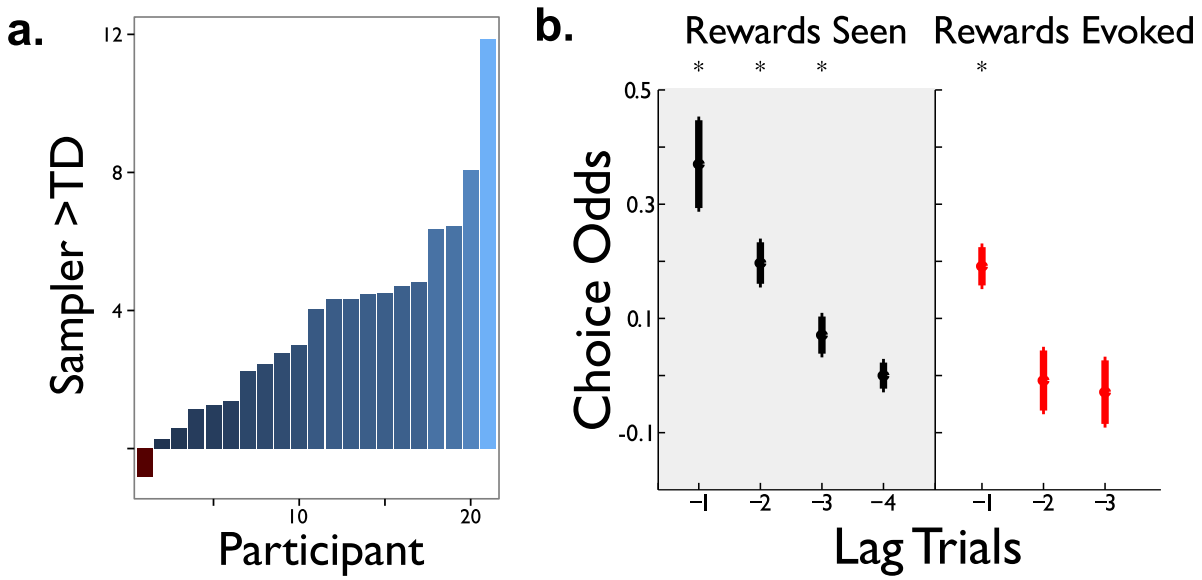

Figure 4: **Ticket bandit results**. a. **Model comparison**. Log Bayes Factors favoring sampling over the TD model. b. **Impact of probes**. As in standard RL models, choices are affected by previously observed rewards (black points). Here, memory probes evoking past decisions (red) also modulate choices on the subsequent choice trial.

this result suggests that a key role in sampling decisions for the episodic memory system and its neural substrates, the hippocampus and medial temporal lobes.

Episode sampling is a potentially powerful mechanism for choice, as it allows generalizing past experience to apply to novel situations in which decision makers can rely on neither learned action values nor structured, semantic knowledge. In these situations, the episodic memory system could use what features partially overlap with past experience to pattern complete and make predictions for entirely new outcomes using elements of old ones [10, 13]. Episodic memories carry associative information besides reward value that can be learned separate from reinforcement and then brought to bear when reward values are introduced. Such an ability to flexibly generalize – as in latent learning and other revaluation tasks – has previously been associated with a "model-based" learning system, distinct from the ("model-free") reward averaging approach [6]. However, although a common assumption is that such behaviors arise from more semantic representations – like spatial maps – learned over multiple experiences [8], it remains possible that samples of individual episodes (e.g. spatial trajectories) might play an analogous role. Thus it remains unclear whether episode sampling like that considered here supports model-based learning in other tasks, or if it represents yet a third way of deciding [18]. Suggestively, we have previously shown that BOLD signal in the hippocampus scales with the difficulty of making these sorts of "latent learning" decisions – consistent with a sampling process that draws on a wider range of memories when associative information is less decisive about the right action to take [12].

A related question is to what extent episode sampling is separate from "model-free" prediction error mechanisms associated with dopamine and its targets in striatum [1, 2]. The finding here that neural RPE signals are captured by expectations formed using associative memories concords with our previous observation that RPE signals during goal-directed decision tasks were better explained by expectations derived from hippocampally-linked associative information [12], and fit with a series of observations that the NAcc RPE signal reflects a mixture of value signals beyond simple model-free TD learning [9, 28]. One possibile unifying explanation for these and the previous findings of model-free RPE is that the systems partly overlap, with samples from memory used to train an average-value representation that guides action selection

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[29]. In this way, rather than computing sampled values separately, direct and sampled experiences might be mixed freely, buffered through a single prediction error signal and net value store [30, 31]. Notably, reward prediction error signals in striatum are influenced by "model-based" and hippocampally linked stimulus-stimulus information [12, 28] and are disrupted when participants successfully form episodic memories for task-irrelevant incidental material [23], suggesting an ongoing, trial-by-trial interaction.

Altogether, although we show here that episode sampling can explain choice behavior and neural signals better than incremental learning alone, normative frameworks support the idea that multiple forms of value learning coexist and their influence on each given choice fluctuates according to momentary demands [6, 32, 33]. This finding concords with those cooperative frameworks by showing that the sampling mechanism may contribute to decision-making continually, not just in situations of low experience or uncertain associative structure (c.f. [18]). A superposition of two recency-weighted processes may be one reason why several studies have observed that actions exhibit a double-exponential form of dependence on past outcomes [25, 34] – with components that have been linked to separate activity in striatum and hippocampus [12, 35]. The results also connect to a known role for episodic memory in ongoing elaborative prospection about future situations [36].

The link between reward-guided choice and episodic memory brings into contact two areas of study with well-developed bodies of computational theory and widespread impacts for cognition more broadly [37], and opens the door for leveraging many other features of episodic memory to explain – and perhaps alter – decisions. Though in the present study we focus exclusively on the direct association between action and reward outcome, as the choice trials contain little other useful information, decisions in the real world take place in contexts laden with associative information that could be of relevance to decisions. It seems reasonable to suspect that adaptive decision-makers can leverage this information to support value estimation.

In a separate study, we replicate and extend the present finding by using a variant of the task presented here to investigate how explicit visual *context* of reminded choice trials impacts sampling-based decisions. We find that choices are in fact also biased by reward information from trials that share such context with the reminded trial [38]. Taken together, these findings open the promise that further studies could investigate how other associative information embedded in episodic memories, such as internal motivational state, might influence decisions.

More broadly, an explicit link between episodic memory and adaptive instrumental control may serve as one rationalization for why it is that organisms encode rich, associatively-dense memories of individual events in the first place [39]. An open question is to what degree the content and persistence of certain memories, rather than others, can be attributed to the potential usefulness of those memories for later decision-making.

Acknowledgements

This work was supported by National Institute of Mental Health predoctoral research fellowship F31MH095501 to A.M.B., National Institute for Neurological Disorders and Stroke R01NS078784 to D.S. and N.D.D., and a Scholar Award from the James S. McDonnell foundation to N.D.D. The authors wish to thank John O'Doherty and Raymond J. Dolan, our collaborators on the study reanalyzed here, and Aubrey Clark-Brown for discovering a bug in an early version of the experiment.

Methods

Experiment 1: Four-armed bandit from Daw et al (2006)

Participants completed a four-choice bandit task, choosing in each of 300 trials between four different slot machines and receiving a payoff (between 0 and 100 points) for their choice. The data analyzed comprise choice timeseries for the 14 fMRI participants reported previously [4], and for behavioral analyses are augmented with 6 additional behavioral-only pilot participants collected for (but not reported with) that study.

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Over the experiment, the mean payoffs for the machines diffused according to independent Gaussian random walks.

For a detailed description of the experimental methods, materials, and previous analyses, see the prior report [4]. This section describes the new analyses we performed.

Analysis: Choice behavior

Two distinct types of models were compared in their effectiveness at explaining each participant's timeseries of choices. The first implemented a temporal-difference learning (TD) approach that kept a running average estimate of action values. The second following a strategy of sampling from prior experiences to estimate these values at the time of decision.

The *standard TD* model maintains a value $Q_t^{TD}(a)$ for each option a, updating it following each choice of a according to the difference, $r - Q_t^{TD}(a)$, between obtained and expected rewards. The amount of update is controlled by a step-size (learning rate) parameter α. Formally, on each timestep we updated the value estimate of each action according to:

$$\begin{aligned} Q_t^{TD}(a_c) &= Q_{t-1}^{TD}(a_c) + \alpha^{TD}[r_t - Q_{t-1}^{TD}(a_c)] \\ Q_t^{TD}(a_u) &= (1-\alpha^{TD})Q_{t-1}^{TD}(a_u) \end{aligned} \tag{1}$$

where r_t is the reward received at trial t, a_c is the chosen action, and a_u is the unchosen action.

Indexing the previous choices of a as $i = 0, 1, 2...t$ choices into the past, this rule can easily be shown [40] at each step to make $Q_t^{TD}(a)$ a weighted average of the rewards previously received for that choice:

$$Q_t^{TD}(a) = \alpha^{TD} * \sum_i (1-\alpha^{TD})^i r_i \tag{2}$$

with weights decaying exponentially (by $1 - \alpha^{TD}$) in the choice lag i.

A similar sort of temporally decaying dependence on past rewards can *also* arise from a strategy of recency-based sampling. Instead of maintaining a running average, the *sampling model* stochastically samples one previous reward for each option with probability given by the same form of weighting:

$$P(Q_t^{sample}(a) = r_i) = \alpha^{sample}(1-\alpha^{sample})^i \tag{3}$$

So that the most recent experience is most likely to be sampled, and previous trials are successively exponentially less likely to be sampled. In our model the aggregate Q^{sample} is computed as the average over the choice probabilities computed using each possible sampled reward, weighted by the likelihood of drawing each trial. (We also considered a version of the model that samples k rewards, with replacement, and averages them to produce $Q_t^{sample}(a)$; this model limits to the standard TD model as $k \to \infty$.)

The models were equated in all aspects other than the functions they used to aggregate past experience. In both models, the resulting decision variables Q_x are linked to choice probabilities according to the standard softmax function with inverse temperature parameter β^x. A third parameter, β^c, modeled any perseverative effect of the previous trial's choice, irrespective of reward received [34, 41]. Choice probabilities were therefore estimated via the combined softmax as:

$$p_t(a = A) = \frac{e^{\beta^c I_t^c + \beta^x Q_t^x(A)}}{\sum_i e^{\beta^c I_t^c + \beta^x Q_t^x(a_i)}} \tag{4}$$

where I_t^c is an indicator function returning 1 if the previous choice (at trial $t-1$) was identical to that on the current trial (t), and 0 otherwise; Q_t^x is the Q value estimated by model x (TD, sampling) at trial t.

Initial conditions. In both models, initial conditions ($Q_0^{TD}(\cdot)$, r_0) were set to the median payout value (50 for Experiment 1 and 0 for Experiment 2). In the sampling model, any sample probability not assigned to trials $i = 0, 1, 2...$ was assigned to r_0. Equivalently, in the TD model, setting a Q_0^{TD} of 50 included that starting value in the running average, with influence decaying according to $1 - \alpha$.

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Model comparison

The likelihood of the sequence of obtained choices (each conditional on the rewards and choices up to that point) is then computed as the product over trials of the softmax probabilities, averaging over all combinations of samples in the sampling model, and optimizing the free parameters (α^x, β^x, β^c, with the number of samples $k = 1$ for the sampling models) separately for each participant and model to maximize this likelihood (minimize the negative log likelihood). We compared the fit of candidate models to the choice data using Bayes factors ([42]; the ratio of posterior probabilities of the model given the data), approximated using the Bayesian Information Criterion (BIC; [43]).

Analysis: fMRI

One participant that is included in the behavioral analysis is excluded from the fMRI analyses due to missing trial onset timing records, leaving 13 fMRI participants considered here.

Region of Interest definition. Based on prior studies, we identified regions of interest (ROIs) for each neural decision variable of interest: Chosen Value (CV), in the ventromedial prefrontal cortex (vmPFC), and Reward Prediction Error (RPE), in the nucleus accumbens (NAcc). The NAcc ROI was defined anatomically according to known structural boundaries using the procedure outlined by Breiter et al and used in our previous studies of reward-guided choice; specifically, defining the nucleus' superior border by a line connecting the most ventral point of the lateral ventricle to the most ventral point of the internal capsule at the level of the putamen [12, 28, 44]. In the case of vmPFC, functional and anatomical boundaries are less well-defined; we defined the current ROI as a 10mm sphere centered on the peak coordinate identified as uniquely responding to "Goal Value" (defined identically to Chosen Value here) by Hare et al. 45.

Regression analysis. Candidate timeseries for each variable of interest (CV and RPE) were generated according to each of the two models: Temporal-Difference learning (TD), and Sampler. The timeseries were first convolved with a hemodynamic response function and then (to account for the hemodynamic lag) we selected the volume two TRs (approximately six and a half seconds) following the relevant event of interest – at decision onset for CV, at outcome receipt for RPE. The resulting timeseries were entered into a simultaneous linear regression, which computed their relative contribution to explaining the timeseries extracted from each of the ROIs. The design matrix also contained twelve participant motion regressors of no interest. The resulting regression weights – indicating the degree to which the candidate variable explained variance in the BOLD signal unaccounted for by the other regressors – were treated as random effects and tested against zero across the population by two-tailed t-test. Finally, the ratio of these weights was also compared against individual participant log Bayes factors contrasting the Sampler model with TD, as described above. This comparison was performed using the robust correlation toolbox, which accounts for outliers that could distort underlying correlations [22], and p-values assessed on the basis of the empirical t statistics returned by this process.

Experiment 2: Ticket bandits

Participants

Thirty individuals (one left-handed; 14 female; ages 18 – 38, mean 23.0) participated in the study. Participants were recruited from the New York University community as well as the surrounding area and gave informed consent in accordance with procedures approved by the New York University Committee on Activities Involving Human participants. All had normal or corrected-to-normal vision. All participants received a fixed fee of \$5, unrelated to performance, for their participation in the experiment. Additional compensation of between \$0 and \$18 depended on their performance on the in-task recognition memory probes (\$0.25 each of 32 probes), two pseudorandomly-selected choice trials (\$5 each), and a post-task recall test. Average total compensation was \$12.

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Exclusion criteria

Data from nine participants were excluded from analysis due to their being unusable, leaving 21 participants analyzed here. Seven participants were excluded due to response biases that indicated they did not attempt to learn the rewards associated with each bandit — specifically, that their choices were more than 90% to either option in either block of the experiment. Two participants were excluded due to below-threshold performance on the in-task memory probes (below the 90% confidence level – or four or more incorrect).

Task design

Participants performed a two-armed restless bandit task — a series of 130 choices between two slot machines — with 32 interspersed memory probe trials, for a total of 162 trials. A compulsory rest period of participant-controlled length was inserted after the 81st trial. The experiment was controlled by a script written in Matlab (Mathworks, Natick, MA, USA), using the Psychophysics Toolbox [46]. Prior to the experiment, participants were given written and verbal instructions as to the types of trials, the buttonpresses required of them, the slot machine payoff probabilities, and the rules for determining final payout. Instructions emphasized that there was no pattern linking the content of the ticket photographs to their dollar value or slot machine. Participants were not told that the memory probe trials should have an effect on their choices, nor was any effect implied. They were, however, told that memorizing the content, source (which slot machine the ticket came from), and dollar value of the tickets would impact their total payout after a post-task memory test.

To aid memory, participants were offered a mnemonic strategy. Specifically, they were told that the photographs could be treated as tickets, which could be placed in one of four imaginary 'pockets': left or right, depending on the slot machine, and front or back, depending on the dollar value. During practice, the keypresses required to advance each trial were instructed in the context of this mnemonic strategy (e.g., "Press 'a' to put the ticket in your left hand, now press 'a' again to put the ticket in your front pocket."). Participants completed two practice trials before beginning the main experiment.

Choice trials

On each choice trial, participants were presented with two slot machines, each on either side of the upper third of the screen (Figure 3a). The slot machines paid out tickets worth either \$5 or -\$5. Participants were instructed to press a key corresponding to the slot machine they felt had the best chance of paying out a winning (+\$5) ticket rather than a losing (-\$5) ticket on that trial. They could choose either the lefthand (key 'a') or righthand (key 'b') slot machine. The slot machines were visually identical except for their color (blue and red), and side of the screen, both of which remained fixed throughout the task. The probability of each machine paying a winning ticket changed independently on each trial according to a diffusing Gaussian random walk with reflecting bounds at 30% and 70% (Figure 3b). At each timestep t, π_t^i — the probability that machine i would pay out a winning ticket — diffused according to: $\pi_{t+1}^i = \pi_t^i + \nu$ for each i. The diffusion noise ν was selected from a zero-mean Gaussian with standard deviation $\sigma_d = 0.1$. The initial payoff probabilities were set to 60% and 40%, with the identity (side, color) of the superior starting bandit psuedorandomly assigned for each participant.

After selection, the unchosen slot machine was covered, and the chosen machine remained alone on the screen for 0.25 seconds. Then, a trial-unique photograph appeared, and remained on the screen until participants again pressed the key corresponding to their chosen slot machine — in the instructions, this corresponded to the mnemonic memory strategy of selecting the left or righthand side 'pockets' for your ticket. When the correct key was pressed, a gray box appeared around the photograph, and remained there for at least 500ms, or additional time up to two seconds depending on when the participants pushed the correct key.

At the end of the timeout, the box disappeared and the dollar value associated with the ticket — either -\$5 or \$5 — was displayed, as a photograph of a \$5 bill, with a green '+' or a red '-' to the left of the bill image. The bill photograph remained on the screen — along with the chosen slot machine, and the trial-unique 'ticket' photograph — until the participant pressed a key corresponding to the value: either 'a' for

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

-$5 or 'b' for $5. In the instructions, this was referred to as putting the ticket in your front or back 'pocket', on the side indicated by the slot machine identity. Once the correct key was pressed, the bill photograph was surrounded by a gray box, and remained on the screen for two seconds. After each trial, a blank screen was displayed for an inter-trial interval of two seconds.

In-task recognition memory probes

Beginning after the tenth choice trial, 32 memory probe trials were interspersed at pseudorandom intervals throughout the task (Figure 4c). Each probe trial consisted of a single photograph and the question: "Is this your ticket? (yes/no)". Twenty six of these photographs were chosen pseudorandomly without replacement from the list of previously seen images; these trials are referred to as *valid* memory probe trials. The remaining six photographs were novel; these trials are referred to as *invalid* memory probe trials. Participants were instructed to press 'a' if they remembered seeing that image, and 'b' if they did not remember seeing that exact image before.

Correct responses — 'yes' for previously seen images, and 'no' for images that were not displayed on a previous bandit trial — were rewarded with $0.25 added to the participant's total payout. This additional reward was indicated by a photograph of a US quarter with a green '+' to the left. Incorrect responses resulted in $0.25 being deducted from the participant's total payout, indicated by a red '-' to the left of an image of a US quarter. Memory probe rewards were displayed for two seconds.

Rewards for memory probes accumulated over the course of the entire task, rather than for randomly selected rounds — so the total payout could be reduced or increased by as much as $8.00. Probe images remained on the screen for up to four seconds — if no answer was entered in that time, the trial was scored as incorrect.

Post-task recall memory probes

Before the experiment began, participants were instructed to remember as many complete bandit-outcome-ticket triplets as they could. Their memory for these triplets was tested in 21 post-task memory probes. Post-task memory probes were drawn only from the images tested during the experiment as valid in-task memory probes. This time, participants were queried as to their recall of each detail associated with the presented probe image: which slot machine it came from, and which dollar value it was associated with. Participants were incentivized to answer correctly by the fact that their total payout in the experiment was predicated on their performance in this memory task. Specifically, payouts for *winning* ($5) tickets were reduced, to $0, if either post-task recall question was answered incorrectly. Payouts for *losing* (-$5) tickets were increased, to -$2.50, if both post-task recall questions were answered correctly (Table 1). (The asymmetry in the final values of the winning and losing tickets was intended to encourage participants to attend to both tasks; if losing tickets could be improved to $0 by correct memory task responses, then participants would have less incentive to track the payout values of the bandits – indeed, simply choosing the same option on every trial would have a net positive EV; similarly, if incorrect memory responses only reduced winning tickets to some positive value, then there would be less incentive to remember the ticket images.)

Analysis

Our analysis of choice behavior in this task addressed the hypothesis that memory probe images affect choices in subsequent bandit trials, by evoking the rewards received by choices on past bandit trials.

Lag kernel analysis

To test this hypothesis, we first examined the effect of past trials — both bandit trials directly experienced, and those evoked by valid memory probes — on choices. Specifically, we entered into a multiple linear regression one regressor for the rewards received on each of the previous ten trials. If a positive reward was received after choosing the right bandit on trial $t - \tau$, this was coded as a 1 in regressor τ, element t. If

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Memory probe outcome/*Ticket value*	*$5*	*-$5*
Correct on both questions	$5	**-$2.5**
Incorrect on one question	**$0**	-$5

Table 1: **Ticket values are modified by performance on a post-task recall memory probe.** After the main slot machine task, 'tickets' paid out by the machines were presented to the participant again. The participant was asked to recall two specific details associated with the ticket: the machine that paid it out (left or right), and the value of the ticket (-$5 or $5). To encourage participants to encode the ticket-machine-value triplet, they were told that the final value of the tickets would depend on both the original value of the ticket and the participant's performance on two post-task recall questions. If they answered either question incorrectly, $5 tickets were modified to be worth $0. If they answered both questions correctly, -$5 tickets were modified to be worth -$2.5. The payout values of each ticket after the memory tests are described in this table. Values altered by the results of the memory tests are higlighted in bold.

a negative reward was received after choosing the right bandit, this was coded as a -1. These values were flipped for lefthand bandits: -1 for positive rewards, and 1 for negative rewards.

Next, we included ten regressors coding rewards from past choices that were evoked by valid memory probes over the past ten trials. Specifically, if trial $t - \tau$ was a valid memory probe that evoked a trial on which the participant received a positive reward for the right bandit, this was coded as a 1 in the regressor τ, element t, and so forth, following the same coding scheme as directly experienced rewards.

We also entered into the regression matching sets of ten regressors each specifying the *identity* of the bandit chosen on each of the previous ten trials: -1 for left, 1 for right. And, similarly, for the *identity* of the bandits evoked by any valid memory probes during the previous ten trials.

Choices of the current bandit were entered as the dependent variable, and coded as -1 (for left) and 1 (for right). The resulting regression weights — indicating the degree to which the current choice was influenced by choices and rewards on a given evoked or directly experienced trial in the past — were tested against zero across the population by two-tailed t-test.

The final regression was thus in the following form:

$$C_t \approx \sum_{i=-1}^{-10} \beta_i^{DR} DR_i + \sum_{i=-1}^{-10} \beta_i^{ER} ER_i + \sum_{i=-1}^{-10} \beta_i^{DI} DI_i + \sum_{i=-1}^{-10} \beta_i^{EI} EI_i \tag{5}$$

where DR is 'directly experienced reward', ER is 'evoked reward', DI is 'directly experienced identity', and EI is 'evoked identity' – each specified over the immediately preceeding ten trials.

Models

We tested a number of specific forms for computational processes by which the bandit trial outcomes evoked by memory probes might affect choices on the current trial.

In the first model, estimated action values were updated only on bandit choice trials; memory probe trials were not considered to have any impact on the values of the slot machines. This model is identical to Equation 2 from Experiment 1, with the exception that the initial value $Q_t^{TD}(\cdot)$ is set to zero, the mean payout on this experiment.

A second class of models incorporated rewards from bandit trials that were evoked by valid memory probes. Specifically, we augmented the models specified in Equations 2 and 3 by adding an additional parameter, α^{evoked}. This parameter allowed the weight given to evoked bandit outcomes to vary, reflecting the fact that the sampling mechanism may itself be stochastic in nature — not every probe trial will successfully trigger a recall of the associated context, even those on which participants exhibit correct recognition memory.

Formally, we distinguish between *direct* and *evoked* rewards r_i according to:

$$P(Q_t^{sample}(a) = r_{i_{direct}}) = \alpha^{sample}(1 - \alpha^{sample})^i \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$P(Q_t^{sample}(a) = r_{i_{evoked}}) \quad = \quad \alpha^{evoked}(1-\alpha^{evoked})^i$$

for the sampling model, and

$$Q_t^{TD}(a) = Q_{t-1}^{TD}(a) + \alpha^{evoked}(r_{i_{evoked}} - Q_{t-1}^{TD}(a)) \tag{7}$$

for a new, augmented TD model.

This formulation was used in computing choice probabilities for all models. The sampling model also had r_0 set to 0, the mean payout on this experiment. These models were otherwise identical to those used in the analysis of Experiment 1.

Another special case nested within the sampling model is that in which the sampling strategy obtains as a value estimation procedure, but does not incorporate rewards received on bandit choice trials evoked by probe images — in other words, the sampling model of Experiment 1. This case is realized when α^{evoked} equals zero.

The likelihood of the sequence of obtained choices (each conditional on the rewards and choices up to that point) is then computed as the product over trials of the softmax probabilities, averaging over all combinations of samples in the sampling model, and optimizing the free parameters (α^x, β^x, β^c, with the number of samples $k = 1$ for the sampling models) separately for each participant and model to maximize this likelihood (minimize the negative log likelihood). We compared the fit of candidate models to the choice data using Bayes factors ([42]; the ratio of posterior probabilities of the model given the data). We approximated the log Bayes factor using the difference between scores assigned to each model via the Laplace approximation to the model evidence [47]. In participants for whom the Laplace approximation was not estimable for any model (due to a non-positive definite value of the Hessian of the likelihood function with respect to parameters), we approximated the log Bayes factor using the difference in likelihoods penalized using the Bayesian Information Criterion (BIC; [43]). Model comparisons are reported both on the log Bayes factors as estimated for each individual, and as aggregated across the population. Parameters were estimated using a maximum *a posteriori* method, accounting for priors over the parameters [47]. The posterior evidence calculations assumed the following prior distributions, chosen to be unbiased over the parameter ranges seen in previous studies [28] (see also Chapter 1), and to roll off smoothly at parametric boundaries: for the learning rate parameters, we employed a prior of $Beta(1.1, 1.1)$; for the softmax temperatures and choice kernel parameters, we employed a prior of $Normal(0, 10)$.

References

[1] Andrew C Barto. Adaptive Critics and the Basal Ganglia. In J C Houk, J L Davis, and D G Beiser, editors, *Models of information processing in the basal ganglia*, pages 215–232. MIT Press, Cambridge, MA, 1995.

[2] W. Schultz, P Read Montague, and Peter Dayan. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599, mar 1997. ISSN 00368075. doi: 10.1126/science.275.5306.1593. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.275.5306.1593`.

[3] Leo P Sugrue, Greg S Corrado, and William T Newsome. Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787, 2004. ISSN 0036-8075. doi: 10.1126/science.1094765.

[4] N D Daw, J P O'Doherty, P Dayan, B Seymour, and R J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879, 2006.

[5] Timothy E J Behrens, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–21, sep 2007. ISSN 1097-6256. doi: 10.1038/nn1954. URL `http://www.ncbi.nlm.nih.gov/pubmed/17676057`.

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[6] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005.

[7] Alan N Hampton, Peter Bossaerts, and John P O'Doherty. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, 105(18):6741–6746, 2008.

[8] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–95, may 2010. ISSN 1097-4199. doi: 10.1016/j.neuron.2010.04.016.

[9] Dylan A. Simon and Nathaniel D. Daw. Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *Journal of Neuroscience*, 31(14):5526–5539, apr 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.4647-10.2011. URL `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4647-10.2011`.

[10] G Elliott Wimmer, Nathaniel D Daw, and Daphna Shohamy. Generalization of value in reinforcement learning by humans. *The European Journal of Neuroscience*, 35(7):1092–104, apr 2012. ISSN 1460-9568. doi: 10.1111/j.1460-9568.2012.08017.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/22487039`.

[11] G.E. Wimmer and D. Shohamy. Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, 338:270–3, 2012.

[12] Aaron M. Bornstein and Nathaniel D. Daw. Cortical and Hippocampal Correlates of Deliberation During Model-Based Decisions for Rewards in Humans. *PLoS Computational Biology*, 9(12):e1003387, dec 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003387. URL `http://dx.plos.org/10.1371/journal.pcbi.1003387`.

[13] Helen C Barron, Raymond J Dolan, and Timothy E J Behrens. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, 16(10):1492–1498, sep 2013. ISSN 1546-1726. doi: 10.1038/nn.3515. URL `http://www.ncbi.nlm.nih.gov/pubmed/24013592`.

[14] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications: Monographs on statistics and applied probability*. CRC Press, 1996.

[15] Dirk Ormoneit. Kernel-Based Reinforcement Learning. *Machine Learning*, 49:161–178, 2002.

[16] Ido Erev and Greg Barron. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112(4):912–31, oct 2005. ISSN 0033-295X. doi: 10.1037/0033-295X.112.4.912. URL `http://www.ncbi.nlm.nih.gov/pubmed/16262473`.

[17] Neil Stewart, Nick Chater, and Gordon D A Brown. Decision by sampling. *Cognitive Psychology*, 53(1):1–26, aug 2006. ISSN 0010-0285. doi: 10.1016/j.cogpsych.2005.10.003. URL `http://www.ncbi.nlm.nih.gov/pubmed/16438947`.

[18] Mate Lengyel and Peter Dayan. Hippocampal Contributions to Control: The Third Way. *Advances in Neural Information Processing Systems*, 20:889–896, 2008.

[19] Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–8, oct 2010. ISSN 1546-1726. doi: 10.1038/nn.2635. URL `http://www.ncbi.nlm.nih.gov/pubmed/20835253`.

[20] G. Giguere and B. C. Love. Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, 110(19):7613–7618, apr 2013. ISSN 0027-8424. doi: 10.1073/pnas.1219674110. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.1219674110`.

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[21] Michael Woodford. Stochastic Choice: An Optimizing Neuroeconomic Model . *American Economic Review*, 104(5):495–500, may 2014. ISSN 0002-8282. doi: 10.1257/aer.104.5.495. URL http://pubs.aeaweb.org/doi/abs/10.1257/aer.104.5.495.

[22] Cyril R. Pernet, Rand Wilcox, and Guillaume A. Rousselet. Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox. *Frontiers in Psychology*, 3(January):1–18, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00606. URL http://journal.frontiersin.org/article/10.3389/fpsyg.2012.00606/abstract.

[23] G Elliott Wimmer, Erin Kendall Braun, Nathaniel D Daw, and Daphna Shohamy. Episodic Memory Encoding Interferes with Reward Learning and Decreases Striatal Prediction Errors. *Journal of Neuroscience*, 34(45):14901–14912, 2014. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.0204-14.2014.

[24] Anne G E Collins and Michael J. Frank. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7):1024–1035, 2012. ISSN 0953816X. doi: 10.1111/j.1460-9568.2011.07980.x.

[25] Brian Lau and Paul W Glimcher. Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys. *Journal of the Experimental Analysis of Behavior*, 84(3):555–579, nov 2005. ISSN 0022-5002. doi: 10.1901/jeab.2005.110-04. URL http://www.pubmedcentral.gov/articlerender.fcgi?artid=1389781.

[26] Jerker Denrell and James G. March. Adaptation as Information Restriction. *Organization Science*, 12 (5):523–538, 2001. ISSN 1047-7039. doi: 10.1287/orsc.12.5.523.10092.

[27] Ido Erev, Eyal Ert, and Eldad Yechiam. Loss Aversion, Diminishing Sensitivity, and the Effect of Experience on Repeated Decisions. *Journal of Behavioral Decision Making*, 21(May):575–597, 2008. doi: 10.1002/bdm.

[28] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and J Raymond. Model-based influences on humans choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011. doi: 10.1016/j.neuron.2011.02.027.Model-based.

[29] Aaron M Bornstein and Nathaniel D Daw. Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, 21(3):374–80, jun 2011. ISSN 1873-6882. doi: 10.1016/j.conb.2011.02.009.

[30] Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, jul 1991. ISSN 01635719. doi: 10.1145/122344.122377. URL http://portal.acm.org/citation.cfm?doid=122344.122377.

[31] Samuel J Gershman, Arthur B Markman, and A Ross Otto. Retrospective revaluation in sequential decision making: a tale of two systems. *Journal of experimental psychology. General*, 143(1):182–94, 2014. ISSN 1939-2222. doi: 10.1037/a0030844. URL http://www.ncbi.nlm.nih.gov/pubmed/23230992.

[32] Mehdi Keramati, Amir Dezfouli, and Payam Piray. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLoS Computational Biology*, 7(5):e1002055, may 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002055. URL http://dx.plos.org/10.1371/journal.pcbi.1002055.

[33] Dylan A Simon and Nathaniel D Daw. Environmental statistics and the trade-off between model-based and TD learning in humans. In J Shawe-Taylor, R S Zemel, P Bartlett, F Pereira, and K Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 127–135, 2011.

[34] Greg S Corrado, Leo P Sugrue, H Sebastian Seung, and William T Newsome. Linear-Nonlinear-Poisson Models of Primate Choice Dynamics. *Journal of the Experimental Analysis of Behavior*, 84(3):581–617, 2005. ISSN 0022-5002. doi: 10.1901/jeab.2005.23-05. URL http://www.pubmedcentral.gov/articlerender.fcgi?artid=1389782.

bioRxiv preprint doi: https://doi.org/10.1101/033910; this version posted December 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[35] Aaron M. Bornstein and Nathaniel D. Daw. Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, 35(7): 1011–1023, apr 2012. ISSN 0953816X. doi: 10.1111/j.1460-9568.2011.07920.x. URL `http://doi.wiley.com/10.1111/j.1460-9568.2011.07920.x`.

[36] Randy L Buckner and Daniel C Carroll. Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2):49–57, 2006. doi: 10.1016/j.tics.2006.11.004.

[37] Daphna Shohamy and Nicholas B Turk-Browne. Mechanisms for widespread hippocampal involvement in cognition. *Journal of Experimental Psychology: General*, 142(4):1159–70, 2013. ISSN 1939-2222. doi: 10.1037/a0034461. URL `http://www.ncbi.nlm.nih.gov/pubmed/24246058`.

[38] Aaron M Bornstein and Kenneth A Norman. Putting value in context: A role for context memory in decisions for reward. *bioRxiv*, 2015. doi: 10.1101/033662.

[39] J R Anderson. A rational analysis of human memory. In HL Roediger III and F I M Craik, editors, *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*, pages 195—-210. Erlbaum, Hillsdale, NJ, 1989.

[40] Hannah M Bayer and Paul W Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–41, jul 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.05.020.

[41] Brian Lau and Paul W Glimcher. Action and outcome encoding in the primate caudate nucleus. *Journal of Neuroscience*, 27(52):14502–14, dec 2007. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.3060-07.2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/18160658`.

[42] R E Kass and A E Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430): 773–795, 1995.

[43] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[44] Hans C Breiter, Randy L Gollub, Robert M Weisskoff, David N Kennedy, Nikos Makris, Joshua D Berke, Julie M Goodman, Howard L Kantor, David R Gastfriend, Jonn P Riorden, R.Thomas Mathew, Bruce R Rosen, and Steven E Hyman. Acute Effects of Cocaine on Human Brain Activity and Emotion. *Neuron*, 19(3):591–611, 1997. ISSN 08966273. doi: 10.1016/S0896-6273(00)80374-8. URL `http://www.cell.com/article/S0896627300803748/fulltext`.

[45] Todd A Hare, John P O'Doherty, Colin F Camerer, Wolfram Schultz, and Antonio Rangel. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, 28(22):5623–30, may 2008. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1309-08.2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18509023`.

[46] D H Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–6, jan 1997. ISSN 0169-1015. URL `http://www.ncbi.nlm.nih.gov/pubmed/9176952`.

[47] David J C Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003. ISBN 9780521642989. doi: 10.2277/0521642981.