Neural Correlates of Value Are Intrinsically History Dependent

Sangil Lee[a*], Caryn Lerman[b], and Joseph W. Kable[a]

[a] Department of Psychology, University of Pennsylvania, Philadelphia, PA, 19104 USA

[b] Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, 90033, USA

[*]Corresponding Author, sangillee3rd@gmail.com

Abstract

A central finding in decision neuroscience is that BOLD activity in several regions, including ventral striatum and ventromedial prefrontal cortex, is correlated with the subjective value of the option being considered, and that BOLD activity in these regions can predict choices out of sample, even at the population-level. Here we show, across two different decision making tasks in a large sample of subjects, that these BOLD value-correlates are intrinsically history dependent. If the subjective value of the previous offer was high, the signal on the current trial will be lower, and vice versa. This kind of history dependency is distinct from previously described adaptation or repetition suppression effects, but instead is of the form predicted by theories of efficient coding such as time-dependent cortical normalization. In terms of practical application, since value-based choice behavior does not exhibit the same history dependence, neural prediction studies may exhibit systematic errors without accounting for history effects. The data-driven, interpretable, whole-brain prediction approach we use to identify history effects also illustrates one way to adjust predictions for neural history dependency.

Keywords: history, value, choice, whole brain, neural predictor, interpretable predictor

Across hundreds of studies, researchers have consistently found neural correlates of subjective value for various commodities (e.g., taste, food, money, drugs, pictures of attractive faces and places, social approval, joy out of others misfortune, viewing of a soccer goal, etc.[1–9]). Meta-analysis has revealed reliable correlates of subjective value across these studies in the ventral striatum (VS) and the ventromedial prefrontal cortex (vmPFC)[10]. Such overlap suggests that there may be a common neural valuation system that operates across all of the tested commodities: a domain-general signal that increases and decreases with subjective value.

Under the assumption that these regions encode subjective value, several studies have shown that activity in VS or vmPFC can predict that individual's future choices[11–14]. Subsequently, and perhaps more surprisingly, several studies have used neural signals from these regions to predict the real-world behavior of other individuals responding to the same stimuli[15–21]. These studies employed a simple approach of using average signals from VS or vmPFC, yet they predicted behavior over and above more traditional non-brain measures[16–21]. For example, Berns and Moore (2012)[18] showed that VS activity of adolescents listening to unknown artists' songs significantly predicted future album sales while their self-report ratings did not; Falk, Berkman & Lieberman (2012)[20] showed that neural activity in vmPFC while watching anti-smoking ads predicted the population-level success of those ads while self-report judgments did not; Venkatraman et al. (2014)[17] showed that activity in VS predicted market-level response to advertising beyond traditional measures; and Genevsky, Yoon, and Knutson (2017)[16] showed that VS measurements from a small sample predicted the market-level outcome of crowdfunding projects while behavioral measures did not.

However, despite these successes in the applied domain, a basic question remains: what quantity, exactly, is being tracked by activity in regions, such as VS and vmPFC, that exhibit

value correlates? Unlike theoretic utility, which is unbounded, neural signals of value are constrained by physiological limits. Hence, in order to efficiently encode values for small stakes decisions (e.g., $1 vs. $2) as well as for large stakes decisions (e.g., 1 million vs. 2 million), the brain needs to adapt its sensitivity to the currently relevant range of values. Previous studies have shown that in blocks with large value options, neural value signals, whether measured with fMRI or single neuron recordings, exhibit lower sensitivity in order to represent a wide range of values, while in blocks with only smaller value options, neural value signals exhibit higher sensitivity to optimally represent the narrow range of values[22–25]. One way to achieve these block-wise changes in sensitivity would be for activity to adjust to recent history in a trial-by-trial manner: specifically, activity would be elevated if the previous trial's value was low and it would be depressed if the previous trial's value was high. Hence, the value signal would not only be positively correlated with the value offered on the current trial, but also negatively correlated with values offered on immediate past trials. Such history dependence is further predicted by theories such as time-dependent cortical normalization[26]. However, though there is some evidence for such trial-by-trial history dependence in single neurons in the orbitofrontal cortex[22], whether history dependence is a general feature of value coding, whether it is present in the BOLD signal, and how the history dependence in different brain regions compares to that in behavior is unknown.

Here, we show that neural correlates of value measured with fMRI are history dependent. Using a large imaging dataset of intertemporal choice and risky choice, we show, separately across both tasks, that neural signals that are positively correlated with the value of the offer on the current trial are also negatively correlated with the recent offer values on past trials. We also show that this history dependency cannot be explained by well-known fMRI adaptation or

repetition suppression effects[27–30]. In contrast to the history dependency in neural activity, we find no evidence that choices are influenced by the recent offer values on past trials. Given this discrepancy, we then build an interpretable whole-brain predictor of choice, as a data-driven method of searching for neural signals of value without history dependency. However, in examining the structure of the whole brain predictor, we find that all positively weighted regions show history dependency and the negatively weighted regions served to subtract out these history effects. In terms of basic neuroscience, these results show that BOLD value correlates are intrinsically history dependent, in a trial-by-trial manner that can account for the range normalization observed in these signals over longer timescales. In terms of practical application, these results show that neural prediction studies that use univariate signals from regions of interest will make systematic errors without accounting for these history effects; the whole-brain multivariate methods we use here provides one way to do so.

Methods

Dataset

We used a dataset from Kable et al. (2017)[31]. The purpose of the study was to examine the effect of commercial cognitive training software on brain activity during decision making tasks. Participants completed two sessions of economic decision-making tasks with 10 weeks in between. In each session, they completed 4 runs each of ITC and RC tasks while being scanned. Each run had 30 binary choices. In ITC, the choices were between a smaller immediate reward of $20 today and a larger later reward (e.g., $30 in 7 days); in RC, they were between a smaller certain reward of $20 and a larger probabilistic reward (e.g., $40 with 60% chance). In both tasks, the smaller amount was always $20, while the larger option varied in amount and delay/probability (**Figure 1b**). The original study reported no difference in decision-making or

brain activity after 10 weeks. Hence, in the current paper, we combine both session 1 and session 2 datasets. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Of the 160 participants that completed session 1, we excluded those with missing runs (n = 6), head movement (any run with >5% of mean image displacements greater than 0.5mm; n = 3), more than 3 missing trials per run for two or more runs (n = 2), or revealed study design (n = 1, one subject expressed awareness of their experimental condition, i.e. cognitive training vs. control). Of the remaining 148, 114 also completed session 2, from which we also excluded those with missing runs (n = 3), head movement (n = 2), or too many missing trials (n = 2). This left us with 148 participants' session 1 data and 107 participants' session 2 data. To ensure the reliability of our choice models, we excluded sessions with extremely one-sided choices (less than 10 trials of either choice type). This removed 20 sessions in ITC and 14 sessions in RC. In total, our final dataset was 235 sessions of ITC data and 241 sessions of RC data across 147 participants and 2 sessions (1 participant's data was completely removed due to one-sided choices in both tasks).

fMRI image acquisition, preprocessing

The images were collected with a Siemens 3T Trio scanner with a 32-channel head coil. T1-weighted anatomical images were acquired using an MPRAGE sequence (T1 = 1100ms, 160 axial slices, 0.9375 x 0.9375 x 1.000 mm, 192 x 256 matrix). T2*-weighted functional images were acquired with an EPI sequence with 3mm isotropic voxels (TR = 3,000ms, TE = 25ms, 53 axial slices, 64 x 64 matrix, 104 volumes). B0 images were collected for unwarping (TR = 1270ms, TE 1 = 5.0ms, TE 2 = 7.46ms). Images were preprocessed with FSL: skull stripped with BET (Brain Extraction Tool), motion corrected with MCFLIRT (Linear Image Restoration Tool

with Motion Correction), spatially smoothed (FWHM 9mm Gaussian Kernel), high pass filtered (cutoff = 104s), and registered to MNI space with FNIRT (FMRIB Non-linear Image Registration Tool).

Behavioral Analyses

Behaviorally, we test if people's choices are affected by the subjective value of previous trials' offers using choice modeling. In each of the two tasks separately, we modeled the logit of each trial's choice probability as a weighted sum of current and previous trials' subjective values, which were simultaneously estimated with a utility function. While researchers often use parametric utility models, there is a myriad of established models for ITC and RC (e.g., exponential [32], hyperbolic [33], generalized hyperbolic [34], quasi-hyperbolic [35]; expected utility theory [36], prospect theory models [37–39], mean variance models[40,41]). Furthermore, using a particular parametric formula assumes that all participants have the same utility function. Hence, instead, we fitted a data-driven utility model via cubic Bezier splines [42], which can approximate any of the established utility models, but can also fit unconventional forms as well. (We note, though, that using other established utility models, such as hyperbolic discounting and power utility functions, does not change our substantive conclusions.)

The model form is shown below. Let Y_t denote the choice at each trial such that $Y_t = 1$ is choosing the larger reward and $Y_t = 0$ is choosing the smaller \$20 at trial t. We fit the following model to ITC and RC data for each session (120 choices) using MLE:

$$logit\big(p(Y_t = 1)\big) = \beta_{scale}\left(\beta_0 + \widehat{SV}_t + \sum_{i=1}^{n} \beta_i \widehat{SV_{t-i}}\right) \tag{1}$$

$$\widehat{SV}_t = Amt_t \cdot CBS(delay_t), \qquad or \qquad \widehat{SV}_t = Amt_t \cdot CBS(probability_t) \tag{2}$$

CBS(delay$_t$) and *CBS(probability$_t$)* shown above denote the discounting function constructed with Cubic Bezier Splines that take delay or probability as input and outputs a discounting factor between 0 and 1. The first *n* trials of each run were removed as they have no lagged variables. Because we estimate the overall scale of the coefficients (β_{scale}), the coefficient of the current subjective value (SV_t) was fixed at 1. The coefficients of the lagged subjective values ($\beta_1 \sim \beta_n$) were constrained between -1 and 1. We fitted two models: lag 4 and lag 0. The lag 4 model was fitted to assess if there was any significant influence of past trials (**Fig. 1c**). Upon concluding that there was not, the lag 0 model was fitted to calculate each trial's subjective value (eq. 2), which were in turn supplied to the fMRI GLM analyses (see below).

fMRI Analyses – GLM

Parallel to the behavioral analysis, we test if neural signals are affected by the subjective value of past trials' offers; separately for both tasks, we regressed each voxel's activity against the intercept, the subjective value (SV) of the offer on the current trial and on each of the previous four trials. To remove their effects from the regression, one nuisance regressor modeled the average activity of the first four trials (trials that have no lagged values). The lagged SV regressors were orthogonalized with respect to the intercept and previous lags (e.g., a lag2 regressor was orthogonalized with respect to the average, lag0, and lag1 regressors). All regressors were time-locked to the trial onset with event duration of 0.1 seconds and convolved with a gamma hemodynamic response function. We used FSL's FLAME (FMRIB's Local Analysis of Mixed Effects) model to obtain an average coefficient of each session for each task. Then we tested for brain regions that have significant activity at the group level by performing permutation tests across all the sessions (235 sessions for ITC, 241 sessions for RC). We used a

threshold of $p = .01$ for each lag (0 ~ 4) so that the Bonferroni familywise error rate would be $p = .05$ for each task.

To thoroughly test whether regions where activity was correlated with the current offer's SV are also influenced by the SV of previous trials' offers, we examined the lagged regressor coefficients in three different ways. First, we created ROIs by choosing voxels from the Bartra et al. meta-analysis [10] that have significant effect of SV in both of our tasks. Using these ROIs for vmPFC and VS, we estimated their average lagged coefficients. Second, we defined spherical ROIs (radius = 2 voxels) around each peak for the effect of the current trial's SV and estimated the average lagged coefficients in these ROIs. Peaks were determined as the local maxima of permutation t-stats within a radius of sqrt(12) voxels (64 ROIs for ITC and 75 ROIs for RC). Last, we examined the raw distribution of lag1 and lag2 coefficients for all voxels where activity was significantly correlated with the current trial's SV at the whole-brain level from the aforementioned permutation tests.

fMRI Analyses – Repetition Suppression

To formally test repetition suppression, we ran 5 GLM analyses each with a different model for repetition suppression. Each GLM included an average activity regressor for all trials, a subjective value regressor, and a distance measure between current and last trial. We employed 5 different distance measures to assess all possible scenarios of repetition suppression: 1) absolute distance between current and previous trial's subjective value, 2) absolute distance between the current and previous trial's reward amount, 3) absolute distance between the current and previous trial's delay/probability, 4) "cityblock" distance between current and last trial's offer in combined attribute space (that is, amount and delay/probability, the additive combination of 2 and 3), and 5) Euclidean distance between current and last trial's offer in combined attribute

space. Each trial's subjective value was calculated using the logit model specified in eq (1) and (2) with lag = 0. Each trial's attribute regressors (i.e., amount, delay/probability) were normalized to [0 1] range (amount was divided by 85, which is the maximum amount; delay was divided by 180 which was the maximum delay). For each combination of task and GLM model, we tested for significant regions of repetition suppression by permutation testing the individual coefficients at the group-level with threshold-free cluster enhancement and two-tailed tests $p < .05$.

fMRI Analyses – MVPA

We combined fMRI data across subjects and the two tasks to build a generalized predictor of choice. First, we estimated the fMRI activity of each single trial via the LS-S (Least Squares-Separate) method[43]. This involved running a separate GLM for each trial with two regressors: one regressor for the trial of interest and one nuisance regressor modeling the average activity of all other trials. The last trial of each run was excluded from this analysis because the scan was often terminated before the onset of blood oxygen level dependent (BOLD) activity of this trial. To downplay the role of noisy estimates, we used the t-stats instead of the raw coefficients. The t-stats were registered to a standard MNI 3mm template before further analysis. Then, we removed voxels that had .4 or less prior probability of being grey matter (FSL's grey matter prior).

We used principle component logistic regression with bootstrap thresholding to create an interpretable whole-brain predictor (PCLR-bootstrap; **Figure 6a**). PCLR-bootstrap starts with principle component decomposition[44] of the predictor matrix, which gives two matrices: the components and their loadings ($\mathbf{X} = \mathbf{TW}$, where $\mathbf{T}$ is component and $\mathbf{W}$ is loadings). The components are then used as predictors in logistic regression ($\text{logit}(\mathbf{Y}=1) = b_0 + \mathbf{T}\hat{\mathbf{b}}$). The

resulting coefficients are then projected back into the brain space by multiplying them with their respective loadings ($\boldsymbol{\beta} = \mathbf{W}\hat{\mathbf{b}}$). This process was repeated 5000 times with bootstrapped trials to obtain the *p*-value of each voxel, which was used to threshold the whole-brain map. The resulting sparse whole-brain coefficient predictor can be conveniently used by simply multiplying it to any new brain image to yield a prediction score (i.e., score = $\mathbf{X}\boldsymbol{\beta}$).

In order to find the best predictive whole-brain map, we performed a grid search of two PCLR-bootstrap tuning parameters while performing cross validation: the number of principle components (400, 600, 800, 1000, 1500, 2000, 2500), and *p*-value threshold ($10^0 \sim 10^{-2}$ in 50 log intervals). For each combination of the two parameters, we performed leave-one-subject-out cross validation across 147 participants. The predictor was trained with 146 participants' data (combined across tasks and sessions) and predicted the left-out participant's data. Out-of-sample prediction performance was measured via the area under the receiver operating characteristic curve (AUROC). The mean of these 147 AUROC scores were used to guide the selection of the best number of principle components and *p*-value. While the best predictive map was constructed with 1500 principal components with threshold at $p = 0.324$, we found that the particular choice of these two parameters were not too important since the parameter combination with the lowest LOOCV performance was only lower by 1% AUROC. For comparison, we also measured the cross-validation AUROC of Bartra et al. ROI predictors which is the mean activity within the ROI.

In order to interpret the whole-brain predictor, we lowered the *p*-value threshold to .01 so that we can focus on the most significant voxels. This led to 44 spatially distinct clusters. One way to understand the whole-brain predictor is to think of it as a sum of these 44 regions' signals, which, in turn, is the weighted sum of each region's voxels' activity according to their assigned

weight maps from the whole-brain predictor. To assess the relationship between these 44 signals, we performed hierarchical clustering analysis using the correlation between the regions' signal as the distance metric. Pairwise correlation between the 44 signals were used as a distance metric (1 – Pearson *r*) and average distance between clusters were used to calculate the hierarchical clustering linkage (MATLAB *linkage* function).

Results

We examined three aspects of history dependencies in valuation. First, in behavioral data, we examined if choices are affected by the subjective value of the offers on previous trials, in addition to the subjective value of the current offer. Then, in the neural domain, we examined if BOLD activity in brain regions where the signal is correlated with the subjective value of the current offer is also correlated with the subjective value of previous trials' offers. As summarized in **Figure 1a**, we found that choices were not affected by the subjective value of the offers on previous trials, but neural activity was affected negatively by them. In light of this potential difference in history-dependence between neural signals and behavior, we built a whole-brain choice predictor to see how neural data can be used to predict behavioral choices.

Participants completed two economic decision-making tasks while being scanned (**Figure 1b**). In ITC, participants made binary choices between a delayed larger monetary outcome, which was shown on the screen, versus an immediate smaller monetary outcome of $20 (fixed across all trials), which was not shown on the screen. In RC, participants made binary choices between a risky larger monetary outcome, which was shown on the screen, versus a certain smaller monetary outcome of $20 (fixed across all trials), which was not shown on the screen. Participants used a button pad to indicate whether they would accept the larger (but

delayed or risky) option on the screen or to reject that offer and receive the smaller (but immediate or certain) $20.

Choice is not affected by the value of previous trials' offers

Behaviorally, we found that choices are not significantly affected by the subjective value of the previous trials' offers. We used logistic regression with lagged subjective values to see if the subjective values of the offers on the past 4 trials have any influence on the current choice. As shown in **Figure 1c**, there is very little influence of past trials' subjective values. To maximize power, we also fitted a simpler model with only the current and immediately previous trials' subjective values. Even then, however, the coefficient of the previous trial's subjective value was not significantly different from zero at the group level and was tightly distributed around zero (sign-rank $z = -0.88$, $p = 0.38$ for ITC; $z = 0.10$, $p = 0.92$ for RC).

Neural correlates of value are affected by the value of previous trials' offers

Neural analyses, on the other hand, revealed many areas where activity is significantly influenced by the subjective value of previous trials' offers (**Figure 2**). We found that activity in many regions is significantly positively correlated with the current trial's subjective value but also negatively correlated with the last trial's subjective value. No negative correlations were found for the current trial's value and no positive correlations were found for the past trials' value. VS and vmPFC were among the regions showing overlapping positive effects of the current trial's value and negative effects of the previous trial's value.

Region of interest analyses further confirm the prevalence of lagged value effects. First, we created ROIs for VS and vmPFC by selecting the voxels that showed significant subjective value effects in both of our tasks (i.e., the conjunction of lag 0 maps from **Figure 2**) within the

regions identified by the Bartra et al. (2013) meta-analysis. We obtained the mean coefficients within each of the two ROIs and then tested if they were significantly different from zero at the group-level (**Figure 3**). In both ROIs and in both tasks, we found that activity was largely negatively correlated with previous trials' values with diminishing effect as the lags increase.

We found similar results across 64 ROIs in ITC and 75 ROIs in RC defined based on the peak positive effects of the current trial's subjective value (**Figure 4**). Across both tasks, an overwhelming majority of the ROIs had negative lag1 and lag2 coefficients, suggesting that regions with a positive effect of subjective value on the current trial tend to have a negative history effect. We also corroborated these results at the voxel level by finding that most of the voxels showing a significant correlation with subjective value on the current trial have negative lag 1 and lag 2 coefficients (**Figure 4**).

Neural history effects are distinct from repetition suppression

Previous fMRI studies have identified repetition suppression effects – reduced activation for repeated stimuli – across a wide variety of different stimuli and brain regions, including in vmPFC. The history dependency we observe in neural value correlates is similar to repetition suppression, in that both involve an influence of the previous trial's stimulus on the current trial's neural response. However, the history dependency described above differs from repetition suppression in that it is independent of the current trial's stimulus, while repetition suppression depends critically on the similarity between the current and previous trial's stimulus. For example, the history dependency described above would predict the activity on the current trial is reduced when the previous trial's offer value is high, regardless of the current trial's offer, while repetition suppression would predict that activity on the current trial is reduced when the previous trial's offer value is high, if the current trial's offer value is also high. Given that our

trials are randomly ordered, the similarity or distance between the current and last trial's offer should be uncorrelated with the past trial's offer, and thus the two effects should be distinguishable in our data.

To show that the history dependency we describe above cannot be accounted for by repetition suppression, we systematically examined five different forms of similarity or distance between the current and previous trials' offers, involving the subjective value of the offers, the attribute values (amount or delay/probability) of the offers, or combinations of the attribute values (assuming additive/"cityblock" or Euclidean distance in two-dimensional attribute space). We found no significant effects of value distance in the brain, and the only distance effects that were consistent across the two tasks were for combined attribute distance in the intraparietal sulcus (**Figure 5**). Critically, none of the distance effects were significant in VS or vmPFC. Even when we used the Bartra ROIs for VS and vmPFC, none of the regions showed significant positive activities for any of the 5 different regressors across both ITC and RC.

Whole-brain neural predictor cancels out the history signal in neural data

Given the discrepancy between strong history effects in neural data but no history effect in behavioral data, we used a data-driven approach to search for any history-independent value signals in the brain. Since choice has no history effects, building a whole-brain predictor of choice should capitalize on any neural signals of value without history effects, if such signals exist. Through a PCLR-bootstrap procedure (**Figure 6**), we created a whole-brain predictor of choice with LOOCV performance (as measured by area under the ROC curve: AUROC) that exceeded that of simple mask-based approaches to prediction.

A hierarchical clustering analysis of the whole-brain predictor (with more stringent thresholding at $p < .01$) revealed two distinct groups of regions that contribute unique signals to the whole-brain predictor (**Figure 7**). We obtained the contribution of each of the 44 contiguous regions to the prediction by calculating the weighted sum of the voxels according to the whole-brain predictor coefficients on every trial. Then the pairwise correlation between these 44 signals across all trials was used to calculate a hierarchical clustering solution. The cophenetic correlation coefficient (ranging from 0 to 1) that measures the quality of the clustering result was very high at 0.98. Furthermore, the first division into two clusters coincides perfectly with whether a region was assigned a negative or a positive weight in the whole-brain predictor. This result could not be achieved if the positive and negative regions were both encoding subjective value but with different signs, as multiplying the negative coefficients to negative subjective value would have made it positively correlated with subjective value. This result suggests that the whole-brain predictor can be thought of as the difference between two signals: the signal carried by the positively weighted regions minus the signal carried by the negatively-weighted regions. VS and vmPFC were in the positively weighted group.

Further examination revealed that the whole-brain predictor works by using the signal in the negatively-weighted group to subtract out the history effects in the signal of the positively-weighted group. To see how the signals are being combined, we examined the average lagged coefficients of the positively-weighted group, the negatively-weighted group, and the two groups combined (**Figure 8**). We found that the signals from the positively-weighted group were quite similar to the signals from the negatively-weighted group; the only difference was relative strength of the subjective value signal compared to the history signal. In the positively-weighted group, the subjective value of the current offer was the strongest signal, whereas in the

negatively-weighted group, the subjective value signal from the current trial was about the same size as the lagged signal from one trial back or even smaller. Due to these differences, when these signals were combined, the whole-brain predictor effectively cancels out the history effects while the subjective value signal from the current trial remains. This allows the whole-brain predictor to match the history dependency of behavior, in which we found no significant effect of the subjective value of past trial's offers on choice. The fact that the whole brain predictor is structured to subtract out history effects is striking; it suggests that such history effects are intrinsic to all brain areas that respond to value, as otherwise the whole brain predictor could have used just those regions without any effects of past history.

Discussion

In the current paper, we present evidence that neural correlates of value are intrinsically history dependent. Separately across two economic decision-making tasks, we find neural signals that are not only positively correlated with the subjective value of the current trial's offer, but also negatively correlated with the subjective values of offers on immediate past trials. On the other hand, choice behavior in both tasks did not show any influence of past trial's offers. We construct a whole-brain predictor of choice, as a data-driven way to search for neural value signals without history effects. However, examination of our whole-brain predictor shows that it does not operate by finding history-free value signals, but rather uses signals from other brain regions to subtract out the history effects intrinsic to value correlates.

Our results have both theoretical and practical implications. On the theoretical side, this is the first evidence for trial-by-trial history dependence in the correlates of value in BOLD fMRI signals. While similar trial-by-trial dynamics have previously been demonstrated in the firing rate of single orbitofrontal neurons[25], our results strongly suggest history dependence is a general

feature of value responses across the brain, at least those observed in the BOLD signal. These trial-by-trial history dependencies could account for the range adaptation observed at longer timescales in the neural value signals measured with either fMRI or single neuron recording[22–25]. Both range adaptation and the trial-by-trial history dependencies observed here may be a result of efficient coding principles, which can favor encoding prediction errors to reduce redundancies in the signal[45], and has recently been proposed to lead to a time-dependent cortical normalization mechanism that impacts value coding in the brain[26,46,47].

Critically, the kind of history dependency we observe for value correlates is distinct from the kind of repetition suppression or similarity/distance effects that have been widely studied in fMRI across different kinds of stimuli and brain regions[27–30]. In fact, we found no significant carryover effects of value distance in the current dataset for either decision task. The distance effects that we did observe in both tasks were for the numerically presented attribute values in the intraparietal sulcus. The anatomical location of these effects is consistent with previous studies reporting repetition suppression in parietal cortex for numerical representations[29].

We did not observe any effects of the subjective value of previous trials' offers on the current trial's choice. This is consistent with a recent proposal that value (or other variables) is encoded according to efficient principles, and that downstream decoding and choice processes assume (and therefore adjust for) such efficient coding[48,49]. However, at least one recent study has shown that people's evaluations adapt to the range of offers across longer timescales (blocks of 90 trials)[46]. Future work should examine whether these differences across studies arise from differences in timescale, decision-making task (choices versus bids), or other factors.

Our whole-brain prediction results also illustrate how building neural classifiers or predictors with an eye towards interpretability can lead to insights into neural mechanisms. Our

PCLR-bootstrap method joins other methods that emphasize interpretability[50,51]. The clustering analysis we performed on the classifier components revealed a central feature of choice-predictive activity – intrinsic history dependence – and how the classifier compensated for this source of "noise".

On the practical side, our results suggest that neural prediction studies that use average activity from ROIs are likely to experience systematic errors unless the history effects are accounted for. In the current study, the whole brain predictor that automatically adjusted for neural history effects predicted choices with significantly higher predictive accuracy out of sample than ROI-based prediction approaches. Using ROI-based approaches, several previous studies have shown that neural measurements from a relatively small number of participants can predict population-level behavior above and beyond traditional behavioral measures such as self-reports[11,15–19]. We believe that using whole-brain predictors, like the one used in the current paper, will increase the predictive power of neural measurements in such applications, in part by canceling out the intrinsic history dependence of neural value signals.

Acknowledgements: This study was funded by grants from the National Cancer Institute R01 CA170297 (JK, CL) and R35 CA197461 (CL)

Author Contributions: Conceptualization, Methodology and Formal Analysis, S.L; Investigation, all authors; Writing - Original Draft, S.L.; Writing Review & Editing, all authors; Supervision J.W.K.; Funding Acquisition C.L., and J.W.K.

Declaration of Interests: The authors declare no competing interests.

Figures

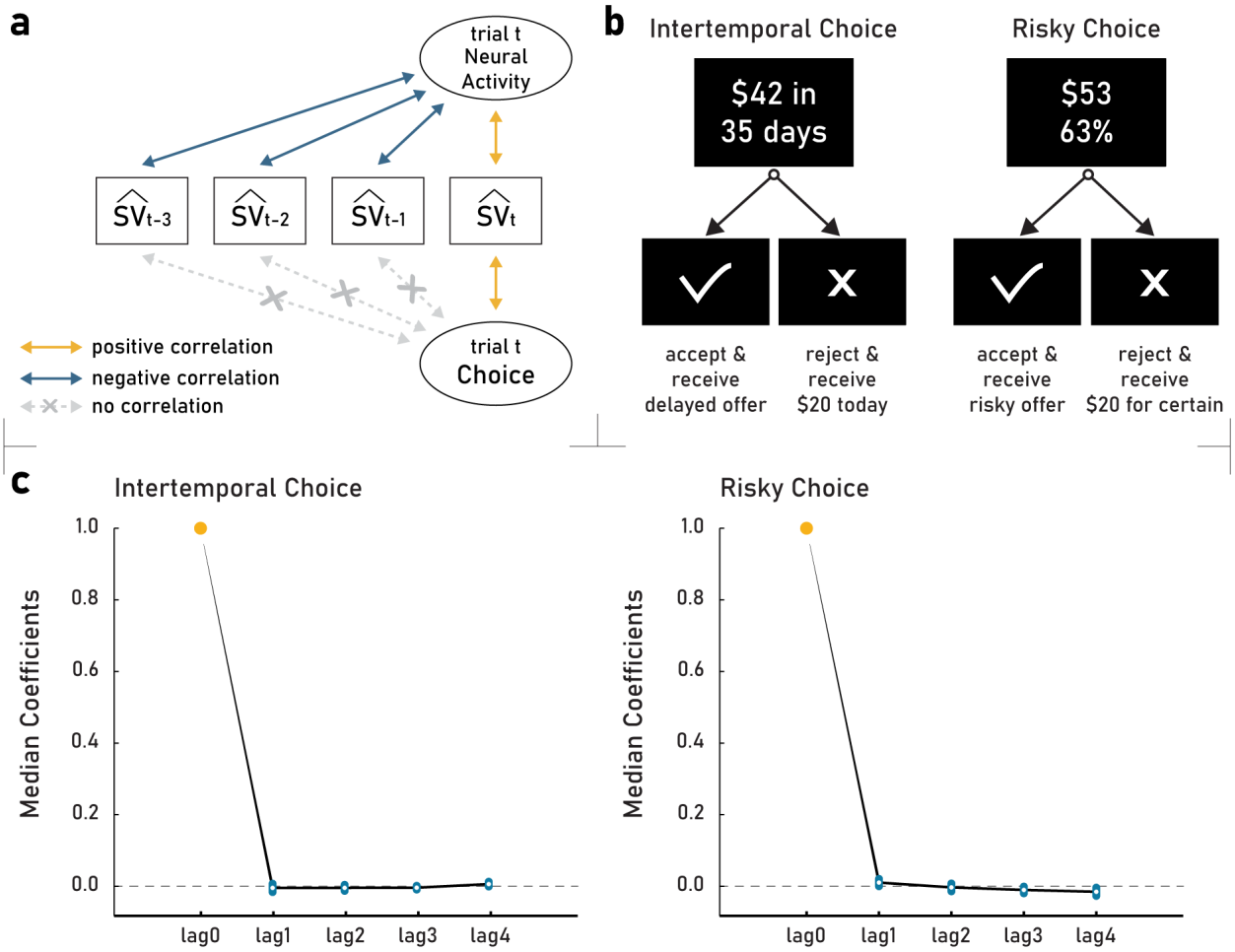

Figure 1. Summaries of regression analyses (a), summary of task (b), and lagged subjective value regression from behavioral data (c). Panel (a) above shows a simplified schematic of the findings in regression analyses. We found that neural activity in value regions was not only positively correlated with subjective value estimates of the current trial's offer but also negatively correlated with subjective value estimates of the past trials' offers. On the other hand, we found no evidence that choice on the current trial was influenced by subjective value of the past trials. Panel (b) below shows a simplified schematic of the two tasks intertemporal choice (ITC) and risky choice (RC). In each task, participants saw a larger monetary option on the screen (delayed/risky) which they could either accept or reject. Panel (c) is the median coefficient estimates from a lag 4 behavioral model (eq. 1) fitted for 235 ITC sessions and 241 RC sessions. The error bars are standard errors of the median calculated via bootstrap. The coefficient of the subjective value (lag0) was fixed at 1, and was not estimated, but is plotted to show relative contribution of each variable. Two-tailed sign-rank tests were performed to test the significance at each lag. Multiple comparison was corrected at the task level using Holm-Bonferroni correction. No lags were significantly different from zero at $p < .05$ level.

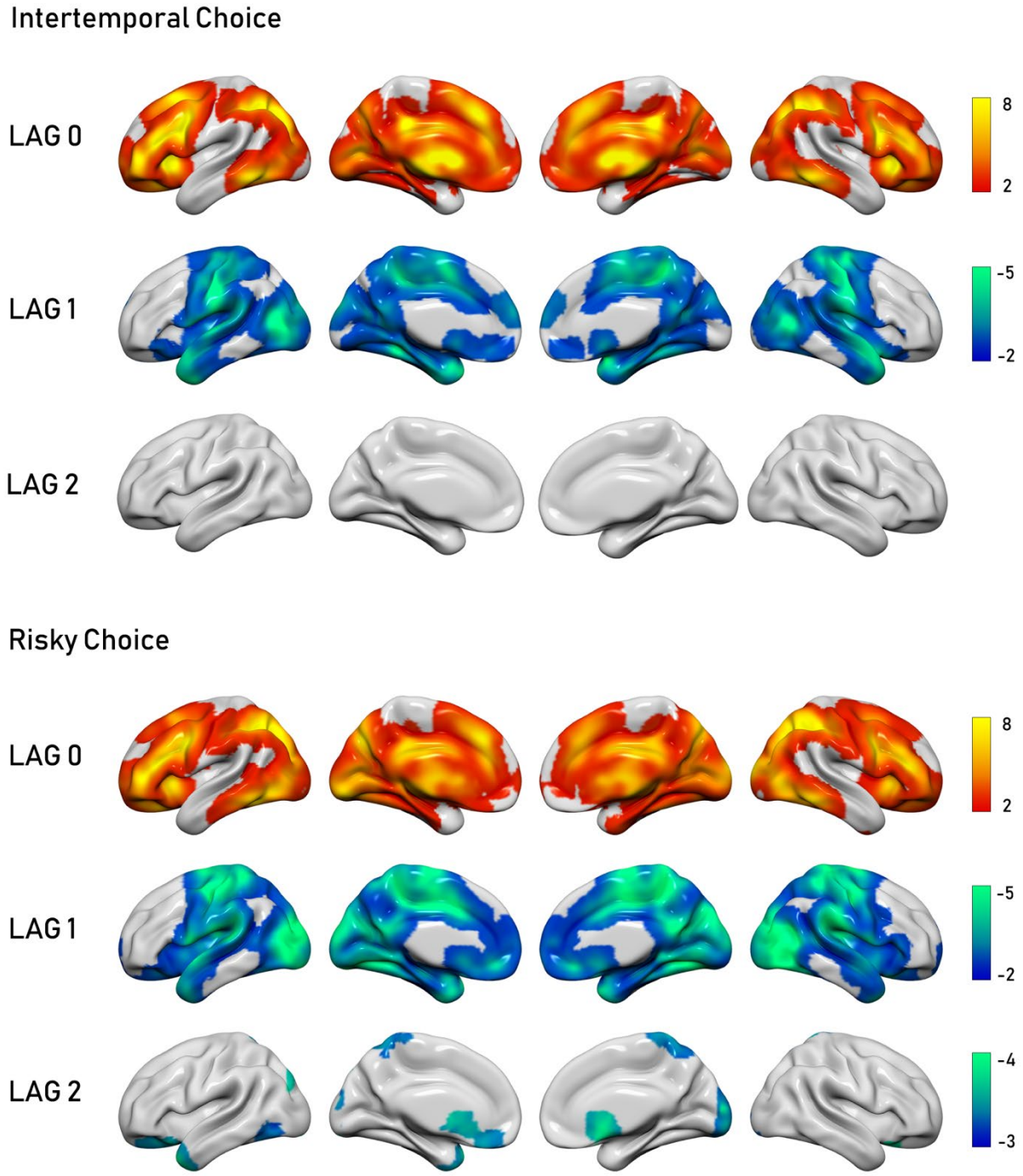

Figure 2. t-statistics from whole-brain permutation tests for GLM lagged coefficients. For each lag (0 ~ 4), we performed two-tailed permutation tests at p = .01 level to find regions that were significantly correlated with each lagged regressor. This corrects for multiple comparisons for each task at p = .05 (Bonferroni correction). No activity was significant beyond lag 2.

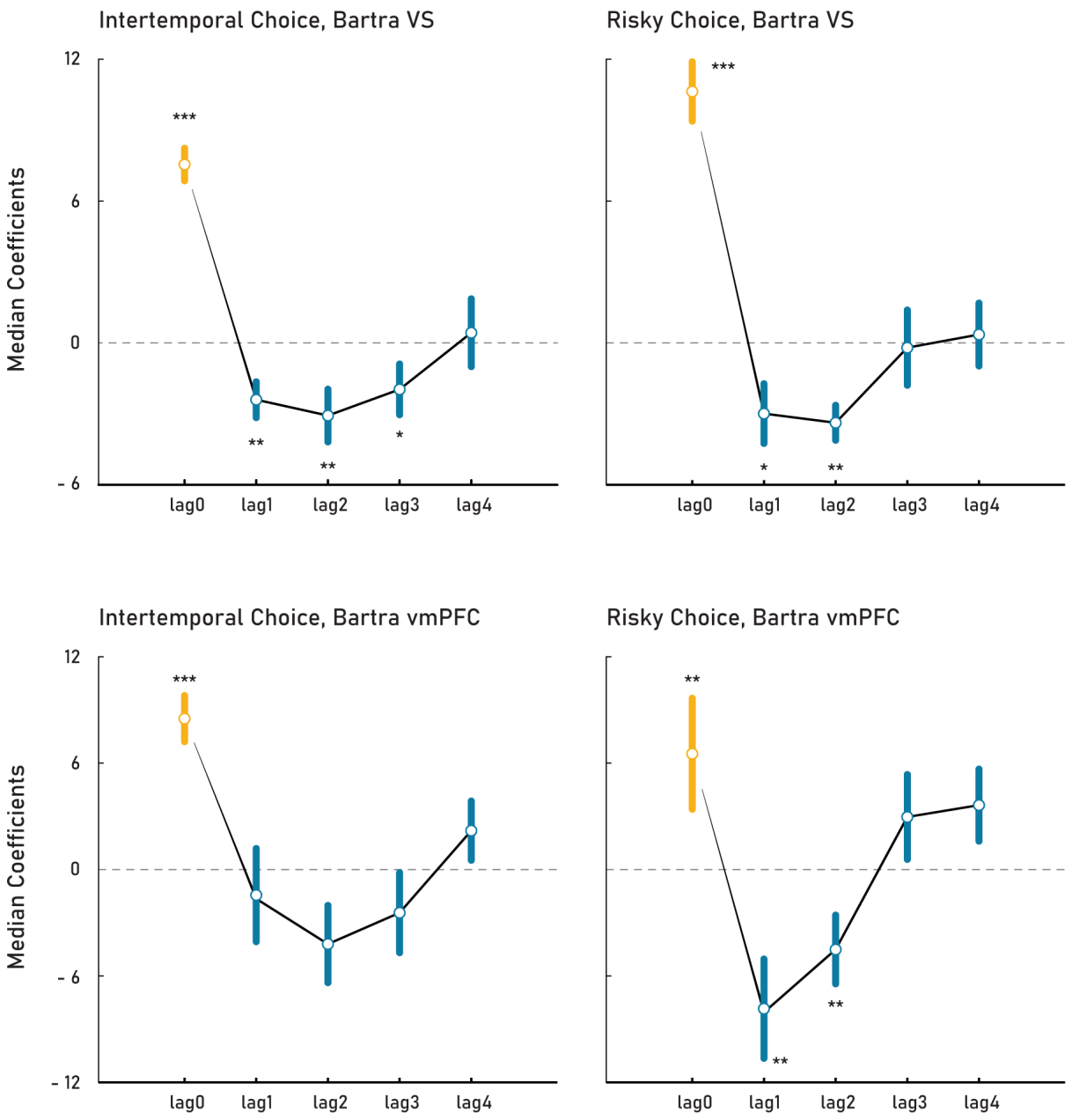

Figure 3. Lagged subjective value coefficients from conjunction mask with Bartra et al. (2013) ROIs. ROIs were created by a three-way conjunction between Bartra et al. masks and permutation tested maps for subjective value in our two tasks. Average coefficient was calculated by taking the median of all coefficients within the ROI. Above plot shows the median of these coefficients across all sessions (235 ITC sessions and 241 RC sessions). The error bars are standard errors of the median calculated via bootstrap. The left panels show the lagged coefficients of ITC and the right panels show that of RC. The top panels show the results from the ventral striatum (VS) ROI and the bottom panels show that from the ventromedial prefrontal cortex (vmPFC) ROI. All significance testing was performed via two-tailed sign-rank tests. * $p < .05$, ** $p < .01$, *** $p < .001$.

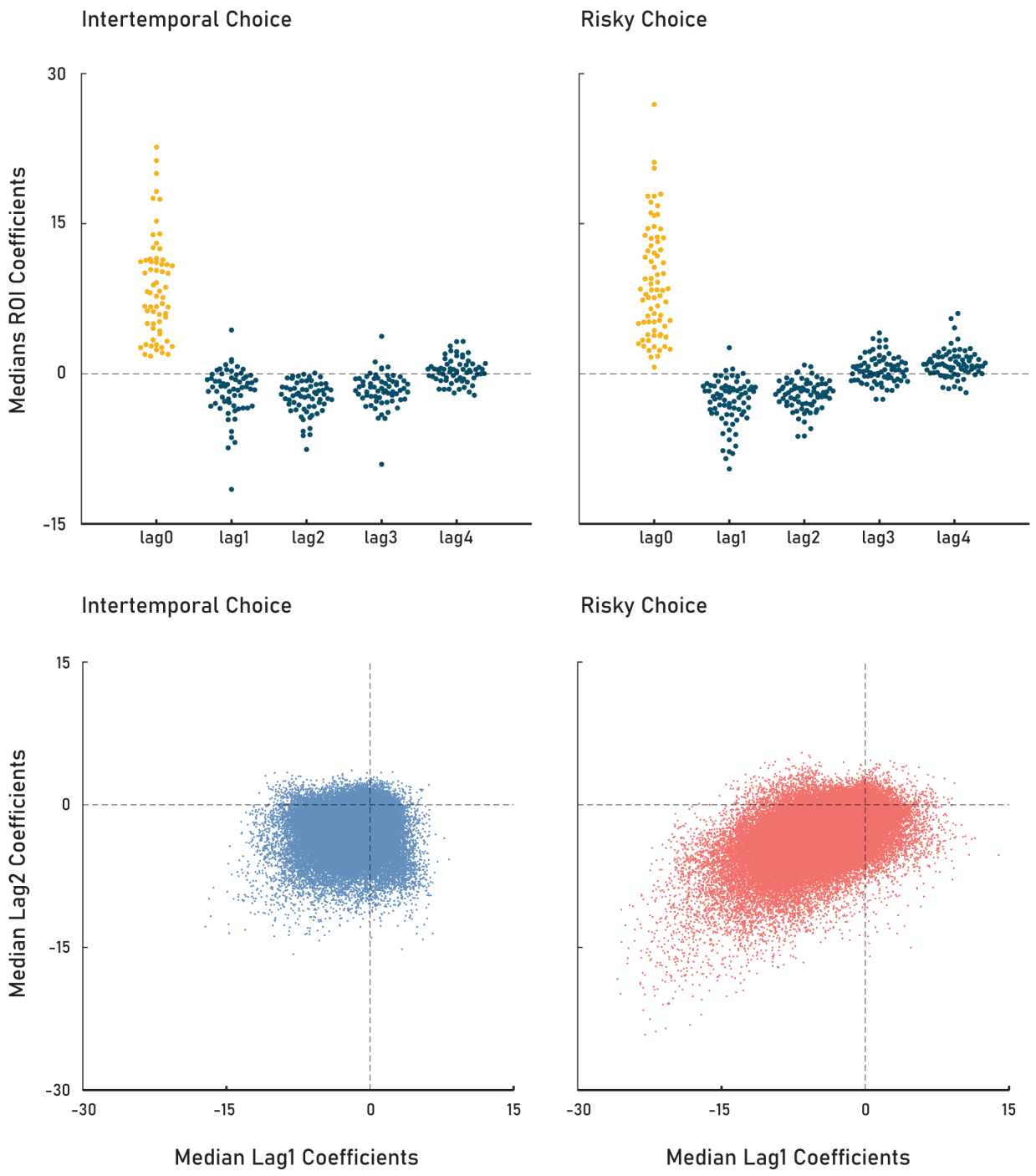

Figure 4. Distribution of lagged coefficients for SV peak ROIs and SV voxels. There were 64 ROIs for ITC and 75 ROIs for RC (each represented as a dot on the upper panel) that were determined by constructing a spherical ROI around the peak voxels of lag 0 t-stat map from figure 3 (i.e., voxels that were shown to be significantly positively correlated with current trial's value at the whole-brain level). The mean coefficient within each ROI was taken as an average measure for each session, and the median of these measures across all sessions are plotted above. Bottom panel is Scatterplot of median lag 1 and lag 2 coefficients for voxels that show significant correlation with subjective value of current (lag0) trial. Each dot is a single voxel whose coordinates are medians of the lag 1 and lag 2 coefficients across all participants.

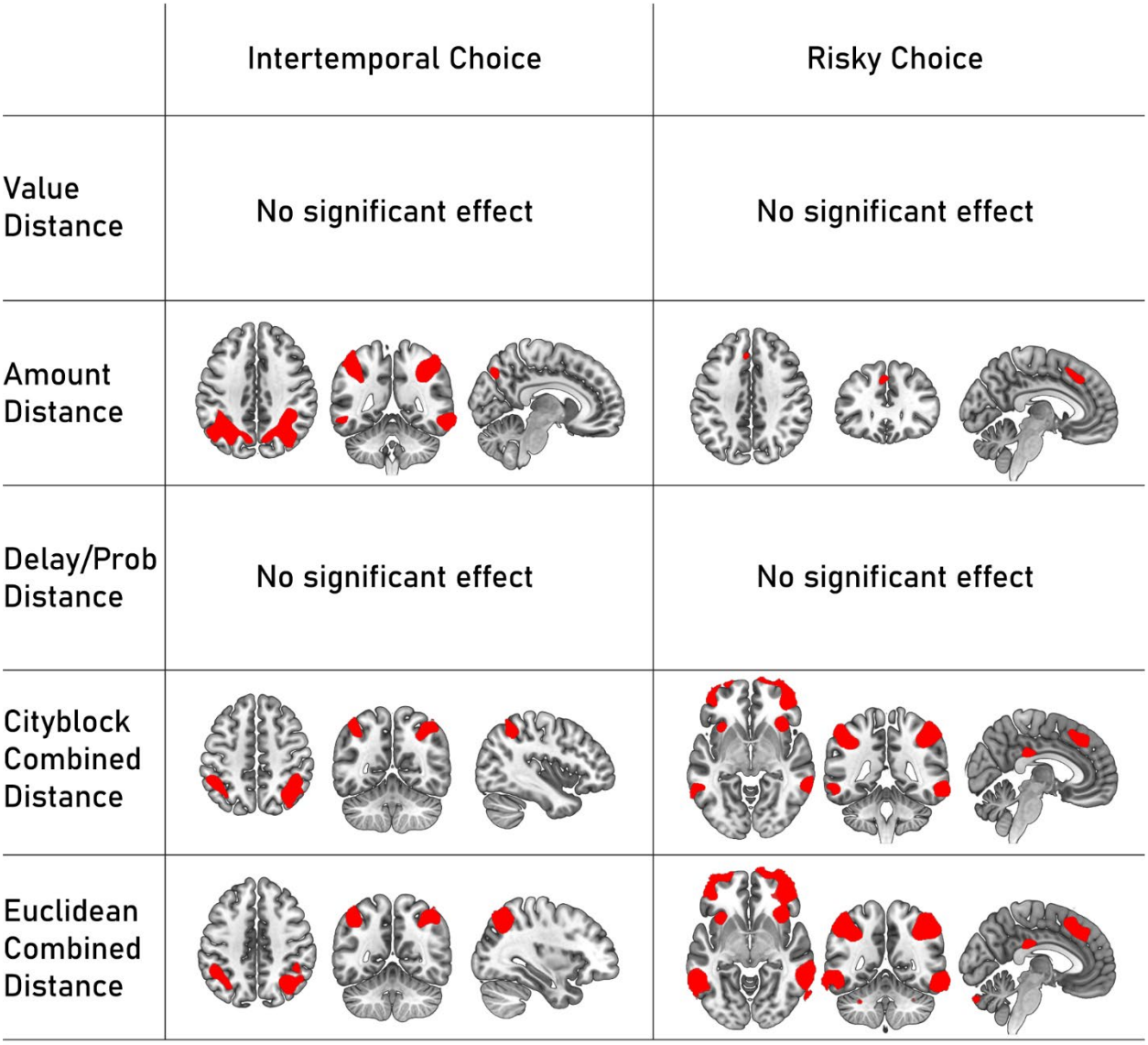

Figure 5. Significant regions for repetition suppression regressors in the brain. Each row corresponds to a different model of repetition suppression. The left column shows significant regions for intertemporal choice dataset and the right column shows significant regions for risky choice dataset. Result for each cell is tested with permutation testing with threshold-free cluster enhancement two-tailed t-tests at $p < .05$. All found results were positive effects. No negative effects were found.

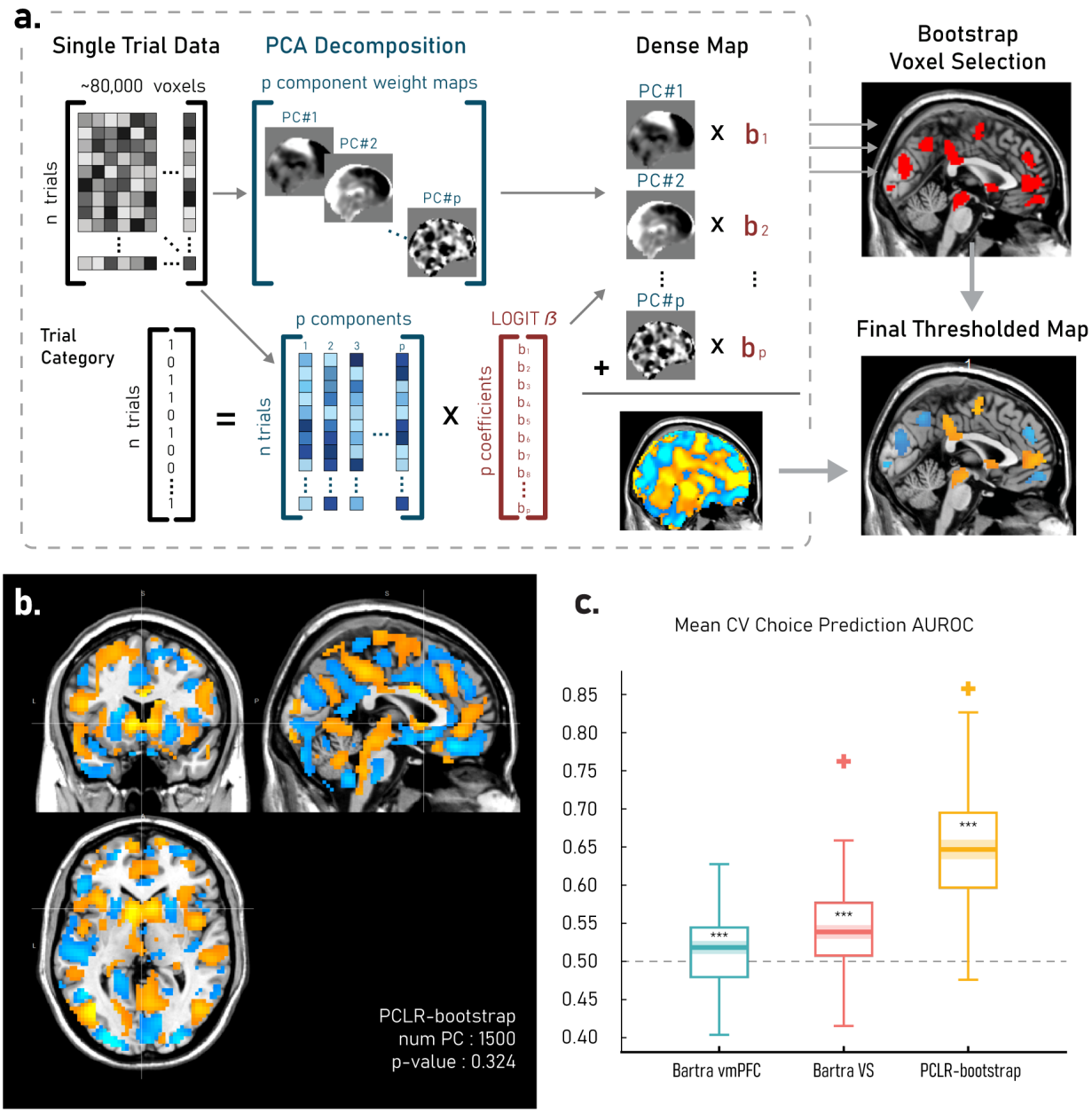

Figure 6. Flowchart of PCLR-bootstrap (a), resulting whole-brain predictor (b), and cross-validation performance (c). Top panel (a) shows the flowchart of PCLR-bootstrap. The original data are composed of single-trial images of the brain and their corresponding behaviors (choices). The single-trial images are then decomposed into principle components, which in turn are used as predictors in a logistic regression. The coefficients of these components, when multiplied with their respective loadings, yield a whole-brain map. The logistic regression is repeated with bootstrapped samples in order to calculate a bootstrap p-value for each voxel, which are then used to threshold the whole-brain prediction map. Panel (b) shows the whole-brain predictor created with PCLR-bootstrap, and panel (c) shows its prediction performance compared with mask-based approaches. All three prediction methods yielded performance significantly higher than 50% (two-tailed sign-rank $p < .001$). The solid line in the middle of the box shows the median, with the shaded thicker line showing 95% confidence interval of the median. The boxes show the interquartile range and the whiskers show the range of data (or 1.5x the interquartile range in case of outliers). The cross mark shows the outliers that are more than 1.5 times the interquartile range away from the nearest interquartile.

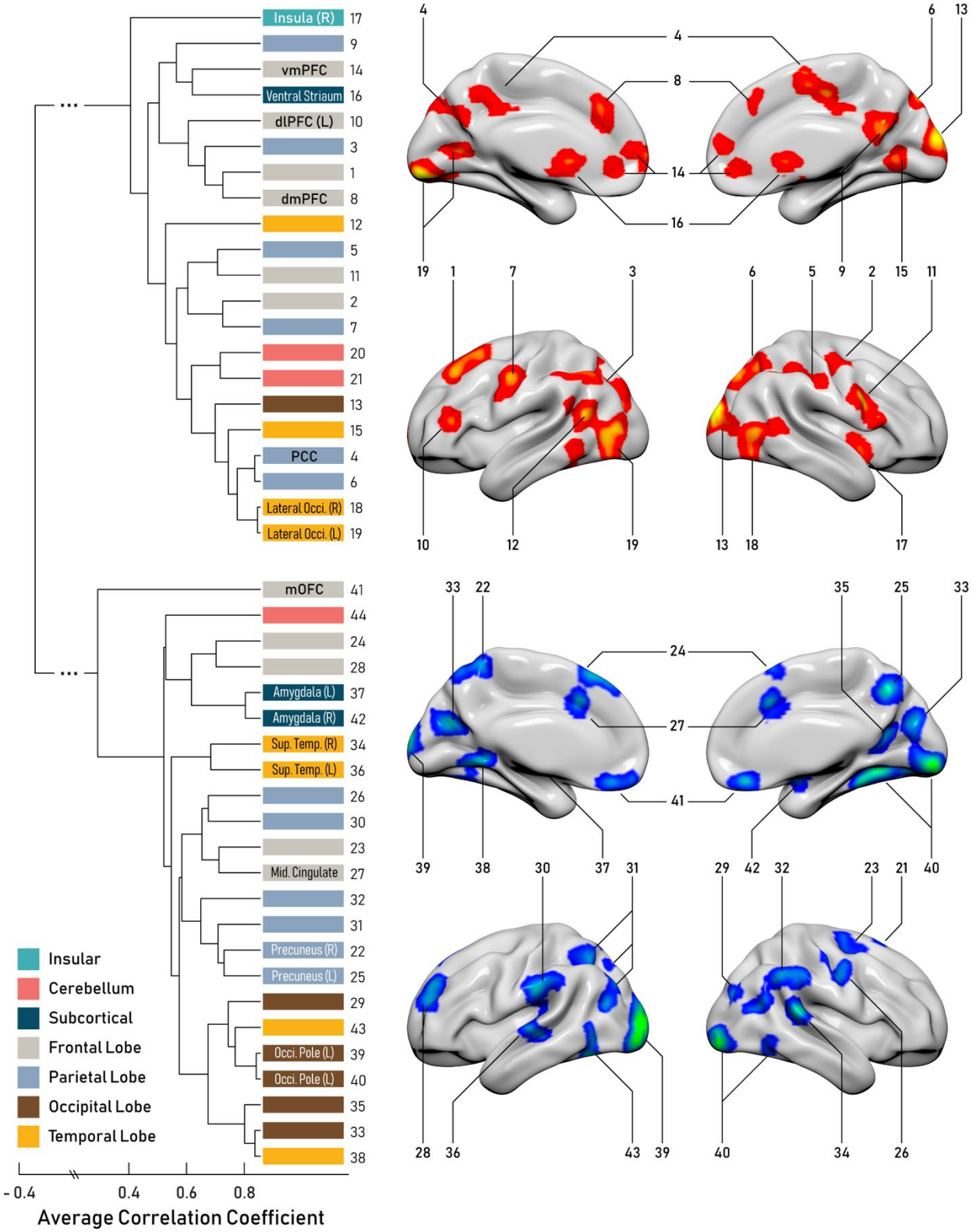

Figure 7. Hierarchical clustering analysis of whole-brain predictor thresholded at p < .01. Shown on the left is a dendrogram that shows the clustering solution and marks the average correlation between groups by the height at which the branch bifurcates. The right shows the corresponding regions in each of the two groups. All positive regions ended up in one group (above) while all negative regions ended up in the other (below).

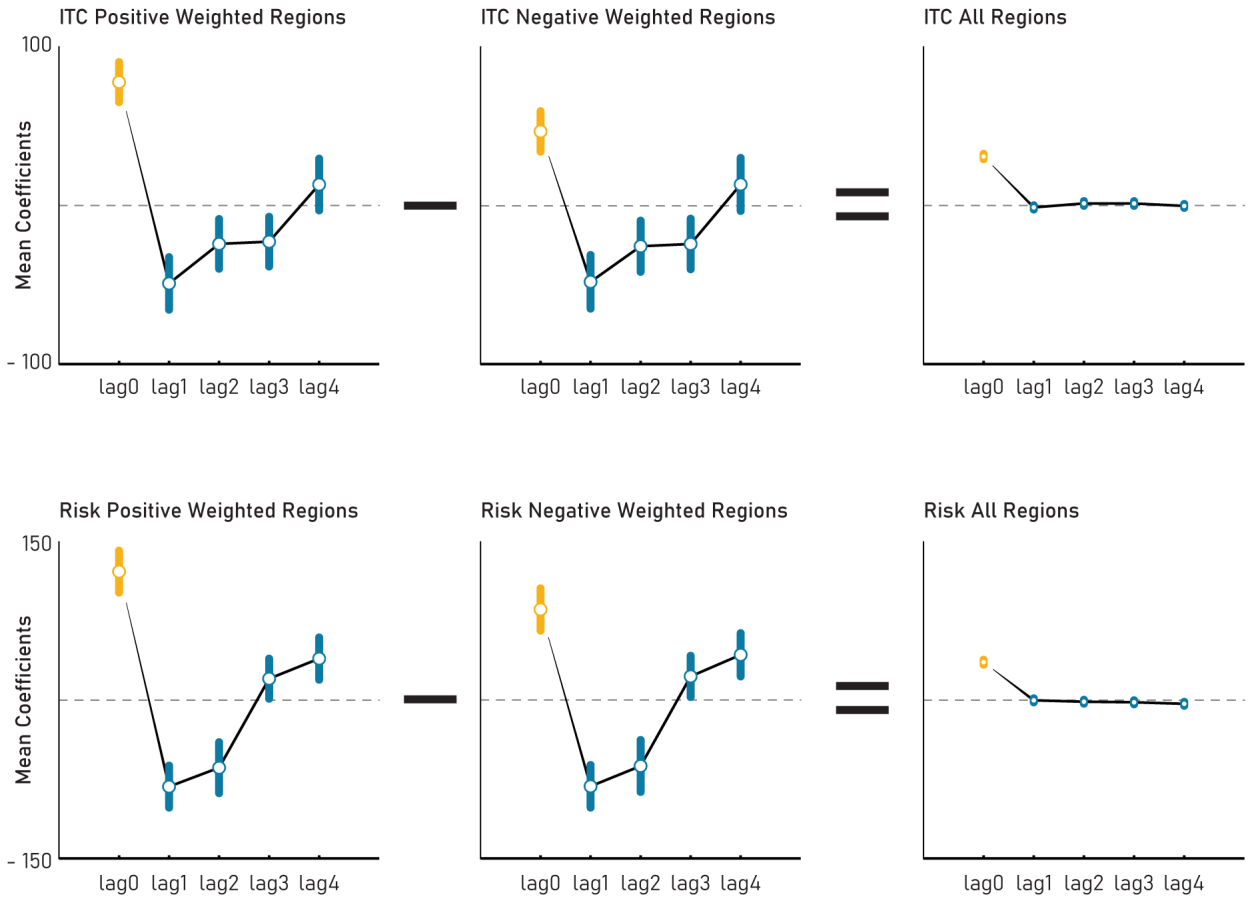

Figure 8. Average history effects of positive regions, negative regions, and all regions of whole-brain predictor. For ITC (above) and RC (below) the whole-brain predictor creates a history-free signal by canceling out the history effect between the regions.

References

1. O'Doherty Deichmann, R, Critchley, HD, Dolan, RJ, J. P. Neural responses during antcipation of a primary taste reward. *Neuron* **33**, 815–826 (2002).

2. Kringelbach, M. L., O'Doherty, J., Rolls, E. T. & Andrews, C. Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cereb. cortex* **13**, 1064–1071 (2003).

3. Knutson, B., Adams, C. M., Fong, G. W. & Hommer, D. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **21**, RC159–RC159 (2001).

4. Breiter, H. C. *et al.* Acute effects of cocaine on human brain activity and emotion. *Neuron* **19**, 591–611 (1997).

5. Cloutier, J., Heatherton, T. F., Whalen, P. J. & Kelley, W. M. Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *J. Cogn. Neurosci.* (2008). doi:10.1162/jocn.2008.20062

6. Pegors, T. K., Kable, J. W., Chatterjee, A. & Epstein, R. A. Common and unique representations in pFC for face and place attractiveness. *J. Cogn. Neurosci.* **27**, 959–973 (2015).

7. Izuma, K., Saito, D. N. & Sadato, N. Processing of the incentive for social approval in the ventral striatum during charitable donation. *J. Cogn. Neurosci.* **22**, 621–631 (2010).

8. Takahashi, H. *et al.* When your gain is my pain and your pain is my gain: Neural correlates of envy and schadenfreude. *Science (80-.).* **323**, 937–939 (2009).

9. McLean, J. *et al.* Localisation of regions of intense pleasure response evoked by soccer goals. *Psychiatry Res. - Neuroimaging* **171**, 33–43 (2009).

10. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).

11. Knutson, B., Scott, R., Wimmer, G. E., Prelec, D. & Loewenstein, G. Neural Predictors of Energy-Efficient Purchases. *Neuron* **53**, 147–156 (2007).

12. Levy, I., Lazzaro, S. C., Rutledge, R. B. & Glimcher, P. W. Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *J. Neurosci.* **31**, 118–125 (2011).

13. Smith, A., Douglas Bernheim, B., Camerer, C. F. & Rangel, A. Neural activity reveals preferences without choices. *Am. Econ. J. Microeconomics* **6**, 1–36 (2014).

14. Tusche, A., Bode, S. & Haynes, J.-D. Neural responses to unattended products predict later consumer choices. *J. Neurosci.* **30**, 8024–8031 (2010).

15. Karmarkar, U. R., Shiv, B. & Knutson, B. Cost Conscious? The Neural and Behavioral Impact of Price Primacy on Decision Making. *J. Mark. Res.* **52**, 467–481 (2015).

16. Genevsky, A., Yoon, C. & Knutson, B. When brain beats behavior: Neuroforecasting crowdfunding outcomes. *J. Neurosci.* 1633–16 (2017). doi:10.1523/JNEUROSCI.1633-16.2017

17. Venkatraman, V. *et al. Predicting Advertising Success Beyond Traditional Measures: New Insights from Neurophysiological Methods and Market Response Modeling*. *Ssrn* **52**,

(American Marketing Association, 2014).

18. Berns, G. S. & Moore, S. E. A neural predictor of cultural popularity. *J. Consum. Psychol.* **22**, 154–160 (2012).

19. Scholz, C. *et al.* A neural model of valuation and information virality. *Proc. Natl. Acad. Sci.* (2017). doi:10.1073/pnas.1615259114

20. Falk, E. B., Berkman, E. T. & Lieberman, M. D. From Neural Responses to Population Behavior: Neural Focus Group Predicts Population-Level Media Effects. *Psychol. Sci.* (2012). doi:10.1177/0956797611434964

21. Genevsky, A. & Knutson, B. Neural affective mechanisms predict market-level microlending. *Psychol. Sci.* (2015). doi:10.1177/0956797615588467

22. Padoa-Schioppa, C. Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. *J. Neurosci.* (2009). doi:10.1523/JNEUROSCI.3751-09.2009

23. Kobayashi, S., Pinto de Carvalho, O. & Schultz, W. Adaptation of Reward Sensitivity in Orbitofrontal Neurons. *J. Neurosci.* (2010). doi:10.1523/JNEUROSCI.4009-09.2010

24. Cai, X. & Padoa-Schioppa, C. Neuronal Encoding of Subjective Value in Dorsal and Ventral Anterior Cingulate Cortex. *J. Neurosci.* (2012). doi:10.1523/jneurosci.3864-11.2012

25. Cox, K. M. & Kable, J. W. BOLD Subjective Value Signals Exhibit Robust Range Adaptation. *J. Neurosci.* **34**, 16533–16543 (2014).

26. Tymula, A. & Glimcher, P. *Expected Subjective Value Theory (ESVT): A Representation of Decision Under Risk and Certainty*. *SSRN* (2016). doi:10.2139/ssrn.2783638

27. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences* (2006). doi:10.1016/j.tics.2005.11.006

28. Barron, H. C., Dolan, R. J. & Behrens, T. E. J. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat. Neurosci.* (2013). doi:10.1038/nn.3515

29. Naccache, L. & Dehaene, S. The priming method: imaging unconscious repetition priming reveals an abstract …. *Cereb. Cortex* (2001).

30. Segaert, K., Weber, K., de Lange, F. P., Petersson, K. M. & Hagoort, P. The suppression of repetition enhancement: A review of fMRI studies. *Neuropsychologia* (2013). doi:10.1016/j.neuropsychologia.2012.11.006

31. Kable, J. W. *et al.* No Effect of Commercial Cognitive Training on Brain Activity, Choice Behavior, or Cognitive Performance. *J. Neurosci.* **37**, 7390–7402 (2017).

32. Samuelson, P. a. Note on Measurement of Utility. *Rev. Econ. Stud.* **4**, 155–161 (1937).

33. Herrnstein, R. J. Self-control as response strength. *Quantif. steady-state Operant Behav.* 3–20 (1981).

34. Green, L., Fry, A. F. & Myerson, J. Research Report Discounting of Delayed Rewards:A Life-Span Comparison. *Pschological Sci.* **5**, 33–36 (1994).

35. Laibson, D. Golden eggs and hyperbolic discounting. *Q. J. Econ.* **112**, 443–478 (1997).

36. Morgenstern, O. & Von Neumann, J. *Theory of games and economic behavior*. (Princeton university press, 1953).

37. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).

38. Goldstein, W. M. & Einhorn, H. J. Expression Theory and the Preference Reversal Phenomena. *Psychol. Rev.* **94**, 236–254 (1987).

39. Prelec, D. The Probability Weighting Function. *Econometrica* **66**, 497 (1998).

40. Markowitz, H. Portfolio selection: efficient diversification of investments. in (Cowles Foundation, 1959).

41. Weber, E. U., Shafir, S. & Blais, A. R. Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychol. Rev.* **111**, 430–445 (2004).

42. Lee, S., Glaze, C. M., Bradlow, E. T. & Kable, J. Agnostic Utility Function Approximation via Cubic Bezier Splines. (2019). doi:10.31234/osf.io/bv2gk

43. Turner, B. A comparison of methods for the use of pattern classification on rapid event-related fMRI data. in *Poster session presented at the Annual Meeting of the Society for Neuroscience, San Diego, CA* (2010).

44. Halko, N., Martinsson, P. G. & Tropp, J. A. Finding Structure with Randomness : Probabilistic Algorithms for Matrix Decompositions. *SIAM Rev.* (2011).

45. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

46. Khaw, M. W., Glimcher, P. W. & Louie, K. Normalized value coding explains dynamic

adaptation in the human valuation process. *Proc. Natl. Acad. Sci.* (2017). doi:10.1073/pnas.1715293114

47. Yamada, H., Louie, K., Tymula, A. & Glimcher, P. W. Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nat. Commun.* (2018). doi:10.1038/s41467-017-02614-w

48. Polanía, R., Woodford, M. & Ruff, C. C. Efficient coding of subjective value. *Nat. Neurosci.* (2019). doi:10.1038/s41593-018-0292-0

49. Wei, X. X. & Stocker, A. A. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* (2015). doi:10.1038/nn.4105

50. Grosenick, L., Greer, S. & Knutson, B. Interpretable classi ers for FMRI improve prediction of purchases. *Analysis* **X**, 1–10 (2008).

51. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* **13**, 1–28 (2015).