Neural Correlates of Context Function in Associative Learning

Samuel Gershman (gershman@fas.harvard.edu), Momchil Tomov, Hayley Dorfman
Department of Psychology and Center for Brain Science,
Harvard University, 52 Oxford St., room 295.05
Cambridge, MA 02138, USA

Keywords: Bayesian modeling; Associative learning; fMRI; Neuroimaging; Context-dependent learning

Introduction

The role of context during learning has been studied extensively and several conflicting views have emerged. Some studies report that the context plays no role during learning (irrelevant context). Another view is that the context modulates cue-outcome associations, thus acting as an "occasion setter" (modulatory context). Yet other studies find that the context simply acts like another punctate cue (additive context). Gershman (2017) reconciled these views in a causal structure learning model. The model captured the behavioral pattern of subjects who learned cue-outcome contingencies that were consistent with a particular context role (irrelevant, modulatory, or additive). In this study, we sought to investigate the neural correlates of context structure learning in human subjects using fMRI.

Materials and Methods

Subjects. We recruited 28 healthy subjects (X female; Y-Z years of age; mean age Z +- SEM) to participate in this study. Eight participants were excluded due to excessive motion or insufficient data.

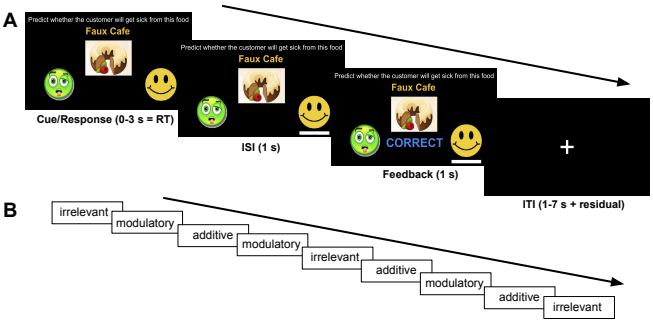

Figure 1: Experimental design. (A) Example timeline of events during a training trial. (B) Example sequence of blocks with the corresponding condition for each block.

Experimental design. We adapted the task used in Gershman (2017) with a mixed within-subjects design consisting of 9 blocks. Each block consisted of 20 training trials followed by 4 test trials. On each training trial, participants were asked to predict whether a particular food (the cue) in a particular restaurant (the context) would cause sickness (the outcome) and were subsequently informed whether their prediction was correct (Figure 1A). On each test trial, participants were asked to make a prediction about an old or a novel cue in an old or a novel context, without receiving any feedback, with each of the 2x2 combinations appearing exactly once. In each block, the cue-outcome contingencies depended on the context in accordance with one of the three causal interpretations (irrelevant, modulatory, or additive) which we refer to as the condition for that block. The nine blocks were divided in three consecutive groups such that each condition appeared in exactly one block in each group (Figure 1B). Each block contained a different set of foods and restaurants that were randomized across blocks.

Simulations. We implemented the model presented in Gershman (2017). The model had two free parameters: the variance σ_w^2 of the Gaussian prior from which the weights are assumed to be drawn; and the inverse temperature β used in the logistic transformation from predictive posterior expectation to choice probability. Intuitively, the former corresponds to the level of uncertainty in the initial estimate of the weights, while the latter reflects the exploration-exploitation tradeoff of the model choices. We fit these parameters using maximum likelihood estimation based on behavioral data obtained from 10 different subjects who performed the same task outside the scanner during a pilot version of the study (data not shown). The fitted values were $\sigma_w^2 = 0.1249$ and $\beta = 2.0064$. All other parameters had the same values as described in Gershman (2017). Each block was simulated independently using the same set of parameters.

Results

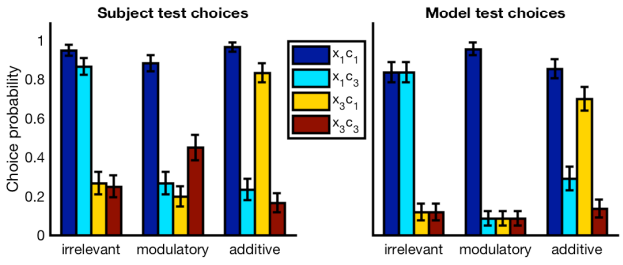

Figure 2: Subject (left) and model (right) performance on the test trials.

Behavioral performance. Test trial choices averaged across blocks are shown in Figure 2. Participants exhibited the same within-subjects behavioral pattern as previously reported using a between-subjects design (Gershman, 2017). The model successfully accounted for participants' choices on both the training and the test trials ($r = 0.7283, p < 0.00001$).

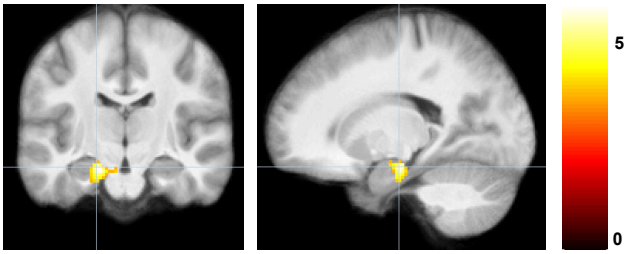

Figure 3: Transient activity related to the additive condition minus the irrelevant condition in left anterior hippocampus (MNI: -18 -16 -18), $t > 3.5518, p < 0.001$, cluster FWE corr. Right: T-value scale.

Imaging data. We had an a priori hypothesis that the hippocampus would be involved in modulating the cue-outcome association when it is influenced by the context. We therefore contrasted BOLD activation at feedback onset between blocks with different conditions. The BOLD signal did not differ significantly between the modulatory and the irrelevant conditions, nor between the modulatory and the additive conditions. The contrast between the additive and the irrelevant conditions showed increased activation in left anterior hippocampus (Figure 3; MNI coordinates of peak voxel: [-18 -16 -18]; T-value: 5.748; extent with $t > 3.5518$: 141; $p < 0.001$; cluster FWE corrected).

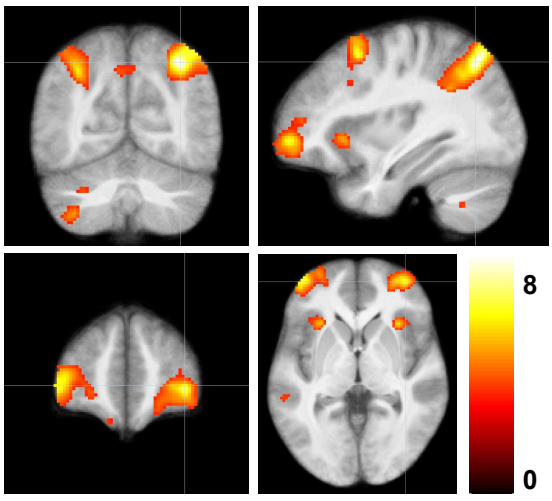

Figure 4: Transient activity tracking the Kullback–Leibler divergence of the posterior over causal structures. Top row: right angular gyrus (MNI: 34 -64 48). Bottom row: right rlPFC (MNI: 48 20 34). $t > 3.5518, p < 0.001$, cluster FWE corr. Bottom right: T-value scale.

In order to measure the neural correlates of structure learning, we computed a contrast with a parametric modulator corresponding to the Kullback–Leibler divergence of the posterior over causal structures as computed by the model at feedback onset on each training trial. The results are shown in Table 1 and Figure 4. We found significant bilateral activation in parietal cortex (the angular gyri), rostrolateral prefrontal cortex (rlPFC), and anterior insula.

Table 1: Brain activation tracking the Kullback–Leibler divergence of the posterior over causal structures. Only cerebral regions with T-value > 5 are shown. All P-values are < 0.001 with cluster FWE correction. Regions were automatically labeled using the AAL2 atlas.

Brain region	Extent	T-value	MNI coord.
Angular gyrus (R)	484	8.638	34 -64 48
Inferior frontal gyrus, opercular part (R)	341	8.378	48 20 34
Middle frontal gyrus (R)	130	7.440	36 56 -2
Middle frontal gyrus (L)	173	7.205	-42 56 2
Middle frontal gyrus (R)	86	6.996	34 12 54
Inferior parietal gyrus (L)	254	6.699	-30 -54 42
Superior parietal gyrus (L)	254	5.566	-34 -72 54
Inferior frontal gyrus, triangular part (L)	173	6.583	-44 20 22
Inferior temporal gyrus (R)	15	6.461	60 -24 -20
Insula (L)	18	6.272	-28 22 -2
Anterior orbital gyrus (R)	8	5.827	20 48 -16

To find out if other regions contain information related to structure learning, we trained a multinomial GLM classifier to predict the block condition based on neural activity from a given region at trial onset. We trained the model on the first 8 blocks and evaluated performance on the last block across all subjects. While performance was at chance level for all regions, the hippocampus showed a significant bias towards predicting the modulatory condition ($p < 0.0001$, one-way ANOVA and post-hoc t-tests) while no bias was observed for other regions (OFC, striatum, and vmPFC; all p's > 0.01). To further investigate this trend, we computed a subject-averaged representation dissimilarity matrix (RDM) using correlation distance of activation patterns from different regions at trial onset. Hippocampal representations of the modulatory condition were more dissimilar from those in the other two conditions ($p < 10^{-6}$, one-way ANOVA and post-hoc t-tests), suggesting that the hippocampus can preferentially distinguish the modulatory causal structure. The same analysis showed that striatum encoded more distinct representations when the context was irrelevant ($p < 10^{-8}$).

Conclusion

These data suggest that different regions encode information about causal structures reflecting the role that context plays during learning. Our results are consistent with previous findings that implicate the hippocampus and the angular gyri in modulating context-dependent associations, with the striatum thought to support a separate stimulus-response learning system.

References

Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic Bulletin and Review*, *24*, 557–565.