Context and Causal Structure in Associative Learning

Momchil Tomov, Hayley Dorfman, Samuel Gershman (gershman@fas.harvard.edu)
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

Context is ubiquitous, yet its role in learning is poorly understood. Some studies suggest that context is irrelevant in some forms of learning (Kaye, Preston, Szabo, Druiff, & Mackintosh, 1987). Other studies suggest that context plays the role of an "occasion setter," modulating cue-outcome associations without itself acquiring associative strength (Swartzentruber, 1995). Yet other studies suggest that context acts like another punctate cue, entering into summation and cue competition with other cues (Grau & Rescorla, 1984). Gershman (2017) reconciled these views within a single model that treats each context role as a distinct causal structure; Bayesian inference can then be used to assign probabilities to each structure based on the training history. The model captured the behavioral pattern of subjects in an associative learning experiment where training history was manipulated to favor particular structures. In this study, we sought to investigate the neural correlates of structure learning in human subjects using fMRI.

Materials and Methods

Subjects. We recruited 20 healthy subjects (10 female; 19-27 years of age; mean age 22 ± 2) to participate in this study. Eight additional participants were excluded due to excessive motion or insufficient data.

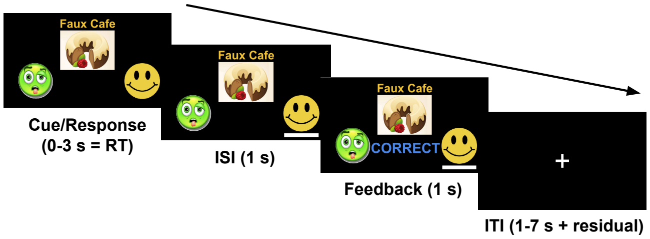

Figure 1: Example timeline of events during a training trial.

Experimental design. We adapted the task used in Gershman (2017) to a 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 1). 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. The training trials within a block were designed to favor a particular causal structure (irrelevant, modulatory, or additive; Table). Each condition was presented 3 times in randomized order.

Condition	Training phase	Test phase
Irrelevant	x_1c_1+	x_1c_1
	x_2c_1-	x_1c_3
	x_1c_2+	x_3c_1
	x_2c_2-	x_3c_3
Modulatory	x_1c_1+	x_1c_1
	x_2c_1-	x_1c_3
	x_1c_2-	x_3c_1
	x_2c_2+	x_3c_3
Additive	x_1c_1+	x_1c_1
	x_2c_1+	x_1c_3
	x_1c_2-	x_3c_1
	x_2c_2-	x_3c_3

Table 1: **Experimental design**. Cues denoted by (x_1,x_2,x_3) and contexts denoted by (c_1,c_2,c_3). Outcome presentation denoted by "$+$" and no outcome denoted by "$-$".

Modeling. 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 choice stochasticity. 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. 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).

Results

Behavioral data. Participants exhibited the same test trial behavioral patterns (Figure 2) in the within-subjects design 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 $(r = 0.7283, p < 0.00001)$.

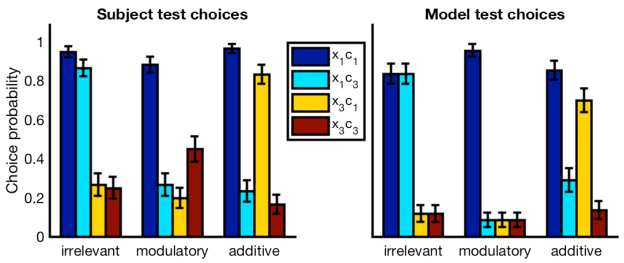

Figure 2: Subject (left) and model (right) performance on the test trials. Each color corresponds to a particular combination of an old (x_1) or new (x_3) cue in an old (c_1) or new (c_3) context.

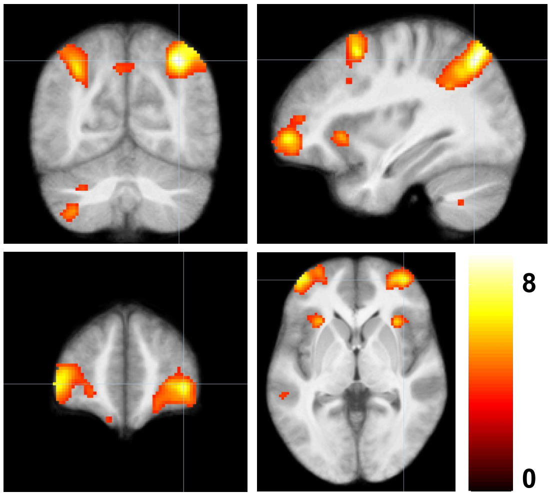

Figure 3: Activity tracking the Kullback–Leibler divergence of the posterior over causal structures. Top row: right angular gyrus (MNI: 34 -64 48). Bottom row: right rlPFC (MNI: 48 20 34). $t > 3.5518, p < 0.001$, cluster FWE corr.

fMRI data. In order to measure the neural correlates of structure learning, we looked for regions that tracked changes in beliefs about the underlying causal structure; specifically, we used the Kullback-Leibler (KL) divergence between the posterior and prior over causal structures as a parametric modulator at feedback onset on each training trial. We found significant bilateral activation in parietal cortex (angular gyri), rostrolateral prefrontal cortex, lateral orbitofrontal cortex, and anterior insula (Figure 3).

To identify regions containing information about causal structure, we trained a multinomial logistic regression 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 ($p < 0.0001$), whereas no significant bias was observed for other regions of interest (orbitofrontal cortex, striatum, and ventromedial prefrontal cortex; 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 from other structures. The same analysis showed that striatum encoded more distinct representations when the context was irrelevant ($p < 10^{-8}$).

Conclusion

These data suggest that several regions encode information about causal structure, reflecting the role that context plays during learning. Some regions exhibit an inductive bias towards particular structures: hippocampus preferentially signals modulatory conditions, whereas striatum preferentially signals irrelevant conditions. 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.

Grau, J. W., & Rescorla, R. A. (1984). Role of context in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, *10*, 324–332.

Kaye, H., Preston, G. C., Szabo, L., Druiff, H., & Mackintosh, N. J. (1987). Context specificity of conditioning and latent inhibition: Evidence for a dissociation of latent inhibition and associative interference. *The Quarterly Journal of Experimental Psychology Section B*, *39*(2), 127-145.

Swartzentruber, D. (1995). Modulatory mechanisms in pavlovian conditioning. *Animal Learning & Behavior*, *23*, 123-143.