Dynamics of neural representations that support generalization under continual learning

Daniel L. Kimmel (dkimmel@columbia.edu)
Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY, USA

Kimberly L. Stachenfeld (ks3316@columbia.edu)
Dept. of Neuroscience & Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY, USA
and Google DeepMind
London, UK

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Dept. of Psychology & Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY, USA

Stefano Fusi (sf2237@columbia.edu)
Dept. of Neuroscience & Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY, USA

C. Daniel Salzman (cds2005@cumc.columbia.edu)
Dept. of Psychiatry & Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY, USA

Daphna Shohamy (ds2619@columbia.edu)
Dept. of Psychology & Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY, USA

Abstract

Abstraction and generalization are essential for flexible decision-making in novel situations. Recent work in humans and monkeys has shown how abstract variables are encoded by the representational geometry of single-neuron population activity. However, these observations are typically made after learning has converged, leaving open the question of how these representations form. To address this question, we developed a factorized model of temporal abstraction that builds on the successor representation. The model disentangles the contributions of different levels of abstract learning—from stimulus-response associations to generalizable task schema—in the form of a *factorized prediction error* that relates the change in relational knowledge to a predicted change in representational geometry on each trial. We fit the model to the behavior of human participants performing a context-dependent decision task during fMRI. The model captured the learning dynamics at multiple timescales, including the increasing contribution of generalization as participants transferred abstracted relational knowledge between novel task instances. In fMRI, BOLD activity in orbitofrontal cortex, hippocampus, and amygdala correlated more with learning attributed to generalization than to the other levels of abstraction. Moreover, the relative dominance of generalization over the other levels increased across task instances in entorhinal cortex, as well as orbitofrontal cortex and hippocampus. Finally, individual variation in the generalization neural signal correlated with behavioral performance on key trials that required relational knowledge. Our findings align with recent proposals for how the brain generalizes abstracted knowledge to current task-relevant states. Our approach offers a computational framework for probing the dynamics of representational geometry under continual abstract learning.

Keywords: abstraction; generalization; continual learning; relational learning; representational geometry

We introduce a computational framework and new experiment with several key features: a) behavioral paradigm permitting continual learning over multiple levels of abstraction; b) computational model quantifying the contribution of each level to relational knowledge; and c) method for relating trial-wise changes in relational knowledge to coincident changes in representational geometry.

Methods

We extended a reversal-learning task from prior studies in monkey and human to include multiple levels of abstract learning (Bernardi et al., 2020; Courellis et al., 2024). During fMRI, healthy human participants (n=41) learned the optimal action and outcome contingencies for a set of stimuli ("local associative learning"). Unbeknownst to participants, the contingencies depended on two latent contexts that alternated in blocks of trials (Fig b). Participants learned and exploited the context-dependent structure: a change in one stimulus was sufficient to update choices for the other stimuli ("cross-context learning"). Across sessions, novel stimuli were used, and participants learned to generalize the task structure to these new instances of the task ("generalization"; Fig c).

To study these multiple levels of learning, we built on the successor representation (SR), a model of temporal abstraction previously applied to planning and navigation and shown to correlate with activity in hippocampus in both animal single neurons and human fMRI (Stachenfeld et al., 2017; Russek et al., 2021; Momennejad et al., 2017; Dayan, 1993; Garvert et al., 2017). The model learned the temporal relationships between behavioral "states" (i.e., unique combinations of stimulus, optimal action, and outcome), which it used to infer the current state (and thereby optimal action) given the previous state. We extended the traditional SR model to include higher levels of abstraction based on recent theory for how associative memory systems learn latent structure (Stachenfeld et al., 2017; Whittington et al., 2020). Critically, our "generalized SR" (gSR) model disentangled the contribution of each level of abstraction in the form of a *factorized* prediction error with separate terms for local associative learning, cross-context learning, and generalization (see "gSR model" for details). When the weights for each factor were fit to behavior, the trial-to-trial magnitude of the corresponding error term revealed the change in a participant's relational knowledge attributed to the respective level of abstraction on any given trial.

Results

Unlike the traditional SR, the gSR model reproduced the multiple timescales of learning (e.g., across contexts and sessions; Fig c) observed in the behavior. Moreover, by interrogating the composition of the prediction error, we found that the basis for learning evolved with experience: participants increasingly relied less on local associative learning and more on generalization, i.e., learning to map the new task instance to the previously-learned abstract structure.

Although the model's internal representation defined a representational geometry analogous to that observed in single neurons (Bernardi et al., 2020; Courellis et al., 2024), we did not test for this geometry directly. Rather, we exploited the factorized prediction error to predict the *change* in representational geometry attributable to each level of abstraction. By testing these trial-wise predictions in fMRI, we probed the *dynamics* of the neural geometry, and obviated the problem of estimating the full geometry on every trial.

Using a single fMRI model, we compared the correlation between BOLD activity and the generalization-based prediction error to the BOLD correlation with the other prediction errors in orbitofrontal cortex, hippocampus, amygdala, and entorhinal cortex—a network of regions implicated in representing relational structure and abstracting shared features across different experiences (Elston & Wallis, 2025; Schapiro et al., 2013; Wikenheiser & Schoenbaum, 2016; Saez et al., 2015;

Garvert et al., 2017). We asked where the contribution of generalization dominated either across all sessions or where it increased from session to session. Overall, BOLD activity correlated more with generalization than the other learning levels in orbitofrontal cortex, hippocampus, and amygdala (Fig d). This aligned with previous single-neuron work showing that neural geometry in the medial temporal lobe and prefrontal cortex represented latent context in a generalizable format in humans and monkeys performing a similar task (Bernardi et al., 2020; Courellis et al., 2024). Crucially, this past work was limited to a static snapshot of the geometry at a point either before or after learning. In contrast, our data relate the trial-to-trial dynamics of learning to the formation of the neural representation *during* learning. In particular, the dominance of generalization's contribution to the BOLD signal, like to the behavior, increased over consecutive task sessions in orbitofrontal cortex, hippocampus, and entorhinal cortex, where grid cells may provide a low-dimensional teaching signal that facilitates transfer of abstract knowledge to novel situations (Whittington et al., 2020; Stachenfeld et al., 2017; Schapiro et al., 2017). Moreover, the component of BOLD activity attributable to generalization correlated with individual variation in behavioral performance on "inference trials", i.e., the small fraction of trials that required relational task knowledge (Fig d).

Our findings support converging evidence for a network of brain regions that learn relational structure and represent it in a generalizable format that can be applied to new instances: entorhinal inputs to amygdala and hippocampus provide relational knowledge abstracted from prior experience (modeled as the generalization term), which are then integrated with current experience to represent the current environment (Stachenfeld et al., 2017; Whittington et al., 2020).

gSR model

Each model term corresponds to a level of abstraction. Local associative learning (α) encodes the local temporal relationships between states using temporal difference learning. Cross-context learning (ω_{self}) extracts *structured* relationships across multiple contexts and longer time spans using low-rank self-regularization. Generalization (ω_{prev}) transfers relational knowledge between environments via low-rank rotation and regularization toward the SR matrix from the prior session.

The 8 unique stimulus-action-outcome combinations define $S = 8$ states. Element (s_{t-1}, s') of the SR matrix $M \in \mathbb{R}^{S\times S}$ represents the expected future occupancy of state s' on trial t given previous state s_{t-1}, from which the agent infers the identity of s'. After the choice, the outcome on trial t serves as a label (correct, incorrect) for self-supervised learning to resolve the current state (i.e, $s' = s_t$). Row $(s_{t-1},:)$ is then updated:

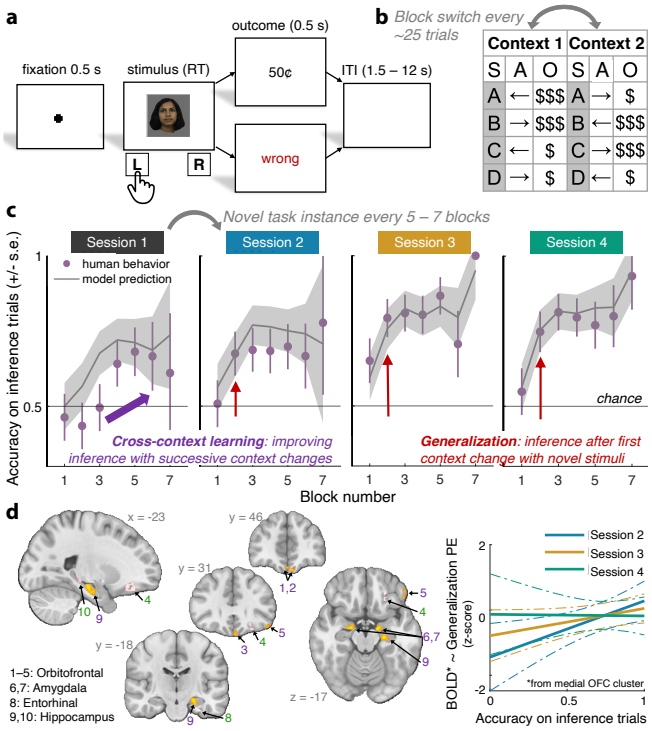
$$\hat{M}_{t+1}(s_{t-1},:) = M_t(s_{t-1},:) + \alpha[\mathbb{1}_{s_t} + \gamma M_t(s_t,:) - M_t(s_{t-1},:)] + \omega_{\text{self}} f^k(M_t)(s_{t-1},:)$$

where $\alpha \in [0,1]$ is the learning rate, $\gamma \in [0,1]$ is the temporal discounting factor (maximal when $\gamma = 0$), and $\omega_{self} \in [0,\infty]$ scales the self-regularization function f (below). The row vector $\mathbb{1}$ has value 1 for the current state s_t and 0 otherwise.

For sessions > 1, the agent generalizes the SR matrix learned from the previous session M_{prev} to the current session by way of the generalization function g (below), which is scaled by $\omega_{\text{prev}} \in [0,\infty]$ and applied to all rows. Combining with the row-updated matrix $\hat{M}_{t+1}$ gives the updated full SR matrix:

$$M_{t+1} = \hat{M}_{t+1} + \omega_{\text{prev}} g^k(M_t, M_{\text{prev}})$$

Functions f and g use the principal components (PCs) of M on trial t, which entorhinal grid cells may provide to amygdala and hippocampus (Stachenfeld et al., 2017). $f^k(M_t)$ self-regularizes M to its top k PCs. $g^k(M_t, M_{\text{prev}})$ assumes that the PCs for any two SR matrices learned on the same meta-structure are equivalent given some rotation R. We estimate R in the space of the top k PCs and regularize M_t toward M_{prev}.



(a) Trial sequence. **(b)** Stimulus (S), action (A), outcome (O) contingency depended on alternating context. **(c)** Choice accuracy on "inference trials" (first encounter with each stimulus after context switch) improved over blocks ("cross-context learning"; *purple arrow*) and sessions with novel stimuli ("generalization"; *red arrows*). **(d)** Clusters where BOLD activity correlated more with generalization prediction error than cross-context (*purple numerals*) or local associative (*green numerals*) learning prediction error either overall (*red-yellow voxels*) or increasing over sessions (*pink voxels)*. Generalization-based component of BOLD activity correlated with behavioral performance on inference trials (far right). (No prior experience to generalize in session 1.)

References

Bernardi, S., et al. (2020, 11). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954–967.

Courellis, H. S., et al. (2024, 8). Abstract representations emerge in human hippocampal neurons during inference. *Nature 2024 632:8026*, *632*(8026), 841–849.

Dayan, P. (1993, 7). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Comp.*, *5*(4), 613–624.

Elston, T. W., & Wallis, J. D. (2025, 1). Context-dependent decision-making in the primate hippocampal–prefrontal circuit. *Nature Neuroscience 2025 28:2*, *28*(2), 374–382. doi: 10.1038/s41593-024-01839-5

Garvert, M. M., et al. (2017, 4). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*, 1–20.

Momennejad, I., et al. (2017, 9). The successor representation in human reinforcement learning. *Nat. Hum. Behav.*, *1*(9), 680–692.

Russek, E. M., et al. (2021, 8). Neural evidence for the successor representation in choice evaluation. *bioRxiv*, 2021.08.29.458114.

Saez, A., et al. (2015). Abstract context representations in primate amygdala and prefrontal cortex. *Neuron*, *87*, 869-881.

Schapiro, A. C., et al. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. doi: 10.1038/nn.3331

Schapiro, A. C., et al. (2017, 1). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. doi: 10.1098/rstb.2016.0049

Stachenfeld, K. L., et al. (2017, 10). The hippocampus as a predictive map. *Nat. Neuro.*, *20*(11), 1643–1653.

Whittington, J. C., et al. (2020, 11). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*(5), 1249–1263.

Wikenheiser, A. M., & Schoenbaum, G. (2016, 8). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nat. Rev. Neuro.*, *17*(8), 513–523.