Disentangling Interpretable Cognitive Variables That Support Human Generalization

Xinyue Zhu
Department of Computer Science
Columbia University
xz3013@columbia.edu

Daniel L. Kimmel
Zuckerman Mind Brain Behavior Institute
Columbia University
dlk2148@cumc.columbia.edu

Abstract

Abstraction and generalization are central to human intelligence. While many algorithms explain how generalization occurs once abstract knowledge is acquired, the mechanisms by which abstract variables are learned remain unclear. One approach is to interrogate computational models that reproduce human behavior. Handcrafted cognitive models are interpretable but rely on strong assumptions about predefined variables. In contrast, recurrent neural networks (RNNs) make fewer assumptions and capture behavior more accurately, yet yield high-dimensional representations of limited interpretability. Here, we use a Disentangled RNN (DisRNN) that uses information bottlenecks to learn a compact set of independent, interpretable latents. Previously, the DisRNN recovered expected mechanisms from simple behaviors. We extend the model to uncover novel mechanisms in a complex task with hidden structure across multiple timescales. The DisRNN was first trained on synthetic data from a handcrafted successor representation (SR) model fit to human behavior, then fine-tuned on data from 41 participants performing the task during fMRI. The model reproduced human learning dynamics across levels of abstraction, including generalizing the task schema to new task instances. Interrogating the model latents revealed a small set of disentangled variables that aligned with the task's abstract structure, providing trial-by-trial estimates of cognitive variables to be tested in neural activity. This framework offers a mechanistic, interpretable account of how humans learn and generalize abstract structure, linking behavioral algorithm to potential neural implementation.

1 Introduction

Flexible human behavior relies on learning general rules or patterns in an abstract format that supports generalization to new situations. However, several central questions remain unknown. What variables are used to represent the abstract structure? How are they learned and deployed from trial to trial? And how is the algorithm implemented in neural activity? One approach to answering these questions is to reproduce human behavior with models whose internal computations are interpretable and can be compared with neural signals.

Handcrafted cognitive models such as the successor representation (SR) infer abstract structure from observable states and outcomes [4, 2, 9, 5]. They assume the cognitive variables and learning rules *a priori*, which affords high interpretability, but by asserting a specific mechanism, risks overlooking an alternative mechanism with greater validity and ability to explain fine-grained behavioral dynamics. Recurrent neural networks (RNNs) learn hidden states directly from data and reproduce behavior with high fidelity, yet their high-dimensional internal activity is hard to interpret or align with neural measurements [6, 7, 10].

We aimed to combine the interpretability of handcrafted models with the expressivity of RNNs. We built on prior work on Disentangled RNNs (DisRNNs) [8], which used information bottlenecks to

yield independent, interpretable latent variables. Applied to simple behaviors, the DisRNN recovered the expected reinforcement learning algorithms, such as Q-learning. Here, we tested the limits of this approach in a richer behavioral domain where the learning mechanism was unknown and multiple algorithms were theoretically possible. In particular, we applied the DisRNN framework to human performance on a decision task where learning unfolded over multiple timescales and levels of abstraction, from local associations to generalization to novel task instances. The ability of the DisRNN to capture the complex dynamics of human learning was unknown. Likewise, multiple learning algorithms had been proposed—each with distinct internal representations—but not disambiguated. By not assuming a particular algorithm, the DisRNN offered a means of generating the intrinsic human strategy. Crucially, because the DisRNN encouraged a compact, disentangled representation, it afforded the opportunity to identify the strategy and produce latent variables that could be directly compared with neural activity.

2 Methods

2.1 Behavioral Task

We adapted a reversal-learning task to engage abstract learning at multiple timescales and levels of abstraction [1, 3]. In each trial, participants viewed one visual stimulus from a fixed set and chose between two responses to obtain a deterministic reward (Fig. 1A). Trials were grouped into blocks, during which the correct stimulus–response–outcome (SRO) contingencies remained stable. Within a block, participants gradually learned the correct response for each stimulus through feedback (*within-block learning*). After several blocks, the SRO contingencies reversed without warning, signaling a change in the latent context (Fig. 1B). To adapt, participants had to infer that the context had switched and update the correct responses for all stimuli based on feedback from only one of them (*cross-block learning*). The experiment consisted of multiple sessions. In each session, a new set of visual stimuli was introduced, but the same abstract task structure was preserved. This allowed participants to transfer previously learned knowledge about the task's structure to novel instances (*cross-session learning*).

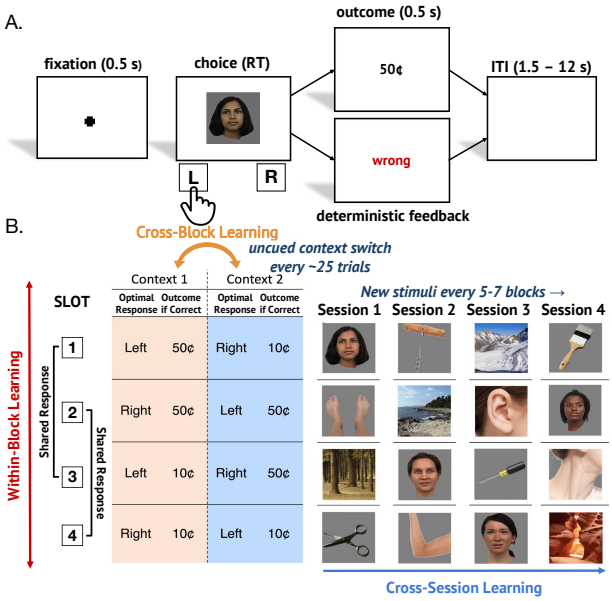

Figure 1: **Human behavioral task with multiple timescales of learning.** **A:** Single trial structure. On each trial, a visual stimulus appeared and participants chose the left or right button to maximize reward given deterministic feedback. **B:** Stimulus-response-outcome contingencies were stable for blocks of trials (*within-block learning*) then switched without cue depending on latent context (*cross-block learning*). Across sessions, participants generalized task structure to novel stimuli (*cross-session learning*). For any session, two stimuli shared the same response, and the other two stimuli shared the opposite response (*brackets*).

2.2 Model Architecture

We used a DisRNN that applies information bottlenecks to disentangle the representations and update rules for task-relevant variables [8]. Unlike a standard RNN, where information is distributed across many hidden units, the DisRNN encourages networks to learn representations limited to a few scalar latents, where each latent corresponds to a single variable, or factor of variation, in the data (Fig. 2).

The bottlenecks penalize the network for using excess information. Each bottleneck serves as a noisy communication channel, parameterized by a learned multiplier m and noise variance σ, such that $\tilde{z}_t \sim \mathcal{N}(mz_t, \sigma)$, where z_t and $\tilde{z}_t$ are the scalar input and output, respectively. When σ is small, information passes through nearly intact; as σ grows, the latent's signal becomes increasingly corrupted. Each bottleneck contributes to the loss, computed as the divergence of the sampling distribution from the unit Gaussian.

Interpretability arises from three bottleneck locations: *update bottlenecks* restrict which external observations and other latents are used to update a given latent, yielding selective and disentangled dependencies; *latent bottlenecks* limit how much information each latent carries forward to the next timestep, encouraging compact temporal memory; and *choice bottlenecks* constrain which latents contribute to the decision network, isolating which computations influence behavior. Each latent is updated by an independent "Update multilayer perceptron" (MLP), and all latents contribute to a separate "Choice MLP" that renders the network's prediction, $\hat{y}$.

The total loss combines the supervised loss from errors in predicting choices with the penalties from all bottlenecks:

$$L_{\text{total}} = L_{\text{softmax}}(y, \hat{y}) + \beta_{\text{update}} L_{\text{update}} + \beta_{\text{latent}} L_{\text{latent}} + \beta_{\text{choice}} L_{\text{choice}}$$

where each bottleneck term is weighted by its own hyperparameter, β.

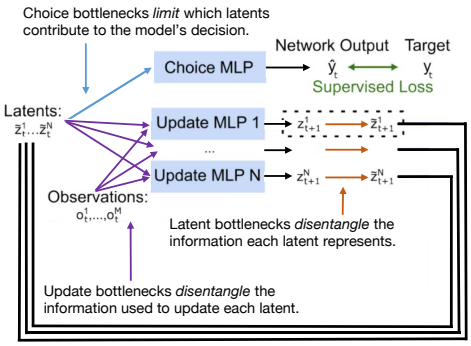

Figure 2: **DisRNN architecture.** The hidden state is defined by a set of scalar latents $z_t^1, \ldots, z_t^N$, each updated by its own residual Update MLP. Update bottlenecks (*purple arrows*) restrict from which external observations and latents each Update MLP receives input. Latent bottlenecks (*orange arrows*) limit what information is carried forward to the next trial. Choice bottlenecks (*blue arrow*) limit from which latents the Choice MLP receives input to produce decision $\hat{y}_t$. Adapted from [8].

We used $N = 10$ scalar latents. Each Update MLP (1 MLP per latent) has 4 hidden layers of 20 units each, and the Choice MLP has 3 layers of 16 units each. On each trial, the network receives binary inputs defining the previous choice, previous outcome, and current stimulus (4 per session * 4 sessions = 16 stimuli). The network outputs the predicted probability of choosing left versus right.

2.3 Training

To train the DisRNN with sufficient data, we first fit a well-validated handcrafted model based on the successor representation (SR) to human behavioral data collected during fMRI. The SR model learned the relationships between observable states, but did not explicitly represent latent variables. Because the human dataset contained limited observations (n = 41 participants), we first trained the DisRNN on synthetic data generated by the fitted SR agent (n = 4000 synthetic agents), which reproduced the human learning dynamics (Fig. 4, blue vs. orange solid curves). The resulting DisRNN model was then fine-tuned on the human behavioral data to more closely capture individual variation in the human behavior and—we presume—more closely align its latent representations with the trial-to-trial neural signals of individual participants.

3 Results

Recovery of Multi-Timescale Learning Dynamics. The DisRNN accurately reproduces choice behavior in both the synthetic SR and veridical human datasets, approaching the theoretical upperbound given the probabilistic nature of the choices and with high cross-validation performance on held-out data (Fig. 3).

We next examine how the DisRNN captures learning dynamics operating at multiple timescales. Within blocks, performance improves across repeated stimulus encounters, reflecting local associative learning of the stimulus-response mappings under fixed contingencies (Fig. 4). The DisRNNs capture these within-block trajectories, and fine-tuning on veridical human data achieves even greater model predictive performance (blue solid vs. dashed lines) than training on a synthetic agent (orange lines).

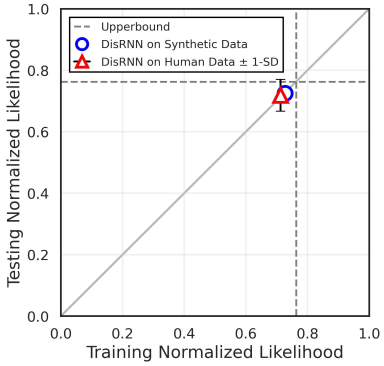

Figure 3: **DisRNN likelihoods.** Scatter plot shows normalized likelihood performance of DisRNN for training (*x-axis*) and testing (*y-axis*) partitions of synthetic SR data (blue) and veridical human data (*red*; average of 10 cross-validation folds). Theoretical upperbounds (*dashed lines*) of maximum possible performance given choice stochasticity.

Each block change probes an agent's knowledge of the task's abstract structure. A single trial of negative feedback after the un-cued change in contingencies is sufficient to infer that the context had switched and to update the responses for the remaining three stimuli. We test for abstract knowledge by examining performance on the first encounter with these stimuli after a context switch, which we refer to as *inference trials*. Overall, accuracy on inference trials is above chance for both human and synthetic agents, indicating agents learn and exploit the task's latent structure.

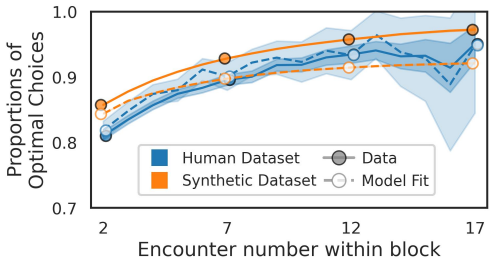

Figure 4: **Choice accuracy on repeated stimulus encounters shows within-block learning.** Separate lines are plotted for data (*filled circles, solid lines*) and DisRNN model fits (*open circles, dashed lines*) from the veridical human (*blue*) and synthetic SR (*orange*) datasets. Shading shows 95% CI.

Across blocks, accuracy on inference trials increases within a session, revealing the dynamics of abstract learning for a fixed set of stimuli (Fig. 5). Across sessions with novel stimuli, above-chance inference accuracy arises even earlier (as early as block 1, consistent with zero-shot learning) and reaches even higher levels, showing transfer of abstract task knowledge to novel task instances. The DisRNN captures these dynamics for both human and synthetic agents. That is, like humans, the DisRNN demonstrates long-timescale learning and a shift to higher forms of abstraction: initially learning the local associations between stimuli and responses, to ultimately acquiring a general schema that supports rapid learning in novel environments by mapping new inputs to a familiar abstracted structure.

Structure of the Learned Representations.
We next examine the DisRNN's internal representations—which the model architecture compresses into a few, disentangled latents—to gain insight into the human learning algorithm. Specifically, 1) the update bottlenecks both disentangle and reveal the inputs from which each latent learns (Fig. 6A, right), and 2) comparing the latent activations to the concurrent task conditions reveals the information each latent represents (Fig. 6B).

Latent 1: shared-response pair. Latent 1 receives input from two of the four stimuli per session, and its activations encode the latent variable "shared-response": stimuli with the same optimal response share the same level of activation (e.g., high), while the complementary stimuli share the opposite level (e.g., low). Strikingly, Latent 1 "listens" to only one stimulus per shared-response pair, which implements a local form of generalization that supports within-block learning: once the correct response is learned for one stimulus in the pair, its partner inherits the code, thereby rapidly updating responses across stimuli, even before they have been encountered in the block.

Latent 2: rapid updating of context. Latent 2 integrates input from Latent 1 (encoding the stimulus) and the previous choice and outcome, which is sufficient to decode the latent variable "context".

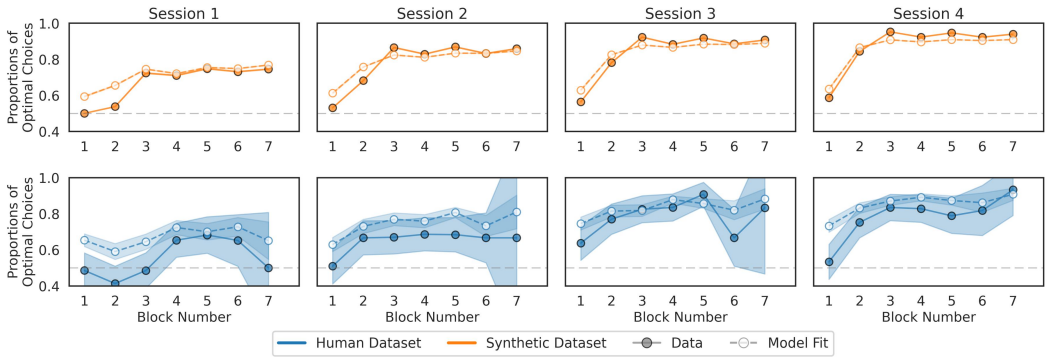

Figure 5: **Choice accuracy on inference trials shows abstract learning and generalization across blocks and sessions.** Average accuracy on inference trials (i.e., first stimulus encounter in block) is shown by block (*x-axis*) and session (*columns*); colors as in Fig. 4. Increasing inference accuracy across blocks and sessions reflects learning of the task's abstract structure and generalization of the structure to novel task instances, respectively. Shading shows 95% confidence intervals. Chance is 0.5 (*horizontal dashed line*)

Indeed, its activation shows rapid, step-like plateaus that flip at block reversals, indicating that it encodes a context variable that updates quickly when surprising outcomes occur. The rapid updating of context, as encoded by Latent 2, enables recovery after reversals and is sufficient to support the high inference-trial accuracy observed in human behavior (Fig. 5).

Latent 3: slow adaptation to changes in context. Latent 3 depends only on Latent 2, and its activation oscillates slowly, with a periodicity that aligns to block transitions. It stabilizes inference by damping the sharp switches in Latent 2 and, importantly, continues to integrate evidence within a block. Even after the context is inferred, uncertainty remains, and Latent 3 accumulates additional evidence with repeated encounters. This slow signal mirrors the gradual within-block improvements in human performance (Fig. 4).

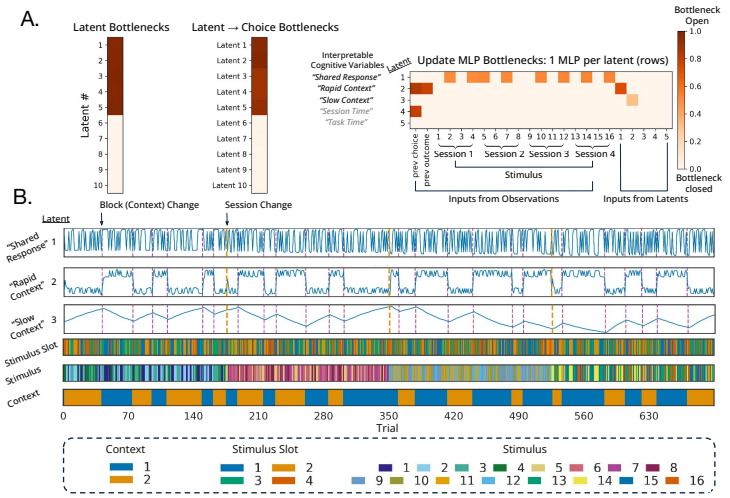

Figure 6: **DisRNN learns compact, interpretable set of task variables. A.** *Latent bottlenecks* (left) restrict learning to subset of 10 candidate latents. *Choice bottlenecks* (middle) restrict decisions to readout of subset of latents. *Update bottlenecks* (right) restrict from which inputs each latent learns. For example, latent 1 receives input from two stimuli with opposite responses from each session. Row labels (*italics*) refer to interpreted variables (see below). **B.** Activation time courses for latents 1–3 are shown, representing key cognitive variables (*row labels*). Colored bars show relevant task variables.

Together, the DisRNN representations suggest a hierarchical learning process: Latent 1 encodes the shared-response pair; Latent 2 integrates stimulus, response and outcome from the prior trial to rapidly infer the latent context; and Latent 3 gradually represents inferred context, permitting evidence accumulation at multiple timescales. The update bottlenecks make this dependency structure explicit: each latent draws only on the information required for its level of abstraction, and once the context and stimulus mapping are represented, the model can use them to produce the choice. Unlike SR models that learn only associations between observable states (e.g., stimulus–response pairings) over which latent states are represented implicitly, the DisRNN reveals a hierarchical representation that supports a distinct algorithm in which latent states (e.g., context) are represented explicitly and from which the observable states are inferred.

4 Conclusion and Future Directions

Our Disentangled RNN (DisRNN) reproduces human abstract learning and generalization at multiple timescales: rapid improvement over repeated encounters within blocks, context inference from sparse outcomes across blocks, and transfer of abstract schema across sessions. It uses information bottlenecks to uncover a small set of interpretable, cognitive variables directly from behavior. By associating the variables with distinct latents and explicit learning rules, the model provides mechanistic insight into the algorithm supporting abstract learning. In so doing, the DisRNN resolves the usual trade-off between model validity and transparency, offering a means to distinguish among competing theories of how abstract knowledge is acquired and applied.

In addition, the DisRNN makes explicit, trial-to-trial predictions about the underlying neural representations. Future work would test these predictions by aligning the latent activations with BOLD activity from concurrent fMRI. More broadly, this approach establishes a framework for discovering highly interpretable cognitive variables, the processes that compute them, and the neural representations that underlie them.

References

[1] Silvia Bernardi, Marcus K Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967, 2020.

[2] Greg Corrado and Kenji Doya. Understanding neural coding through the model-based analysis of decision making. *Journal of Neuroscience*, 27(31):8178–8180, 2007.

[3] Hristos S Courellis, Juri Minxha, Araceli R Cardenas, Daniel L Kimmel, Chrystal M Reed, Taufik A Valiante, C Daniel Salzman, Adam N Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026):841–849, 2024.

[4] Nathaniel D Daw et al. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, 23(1):3–38, 2011.

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

[6] Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, and Bernard W Balleine. Models that learn how humans learn: The case of decision-making and its disorders. *PLoS computational biology*, 15(6):e1006903, 2019.

[7] Matan Fintz, Margarita Osadchy, and Uri Hertz. Using deep learning to predict human decisions and using cognitive models to explain deep learning models. *Scientific reports*, 12(1):4736, 2022.

[8] Kevin Miller, Maria Eckstein, Matt Botvinick, and Zeb Kurth-Nelson. Cognitive model discovery via disentangled rnns. *Advances in Neural Information Processing Systems*, 36:61377–61394, 2023.

[9] JOHN P O'DOHERTY, Alan Hampton, and Hackjin Kim. Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of sciences*, 1104(1):35–53, 2007.

[10] Mingyu Song, Yael Niv, and Mingbo Cai. Using recurrent neural networks to understand human reward learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.