Interpretable Hybrid Neural-Cognitive Models Discover Cognitive Strategies Underlying Flexible Reversal Learning

Chonghao Cai* ETH Zurich
Liyuan Li* ETH Zurich
Yifei Cao* UCLA
Maria K. Eckstein Google DeepMind

Abstract

Flexible learning in dynamic environments is classically studied with reversal learning tasks, but existing reinforcement learning models often fail to capture the full richness of behavior. Here we used HybridRNNs to model human and primate reversal learning. Among several variants, the Context-ANN, a model variant that replaced linear value belief updating rule with neural network and additional contextual information input, achieved the highest predictive accuracy, closely matching trial-by-trial adaptation to reversals. Analyses of its internal dynamics revealed a distinctive, context-dependent value-updating strategy with non-linear attractor structures, providing interpretable insights into how flexible learning is implemented. These results show that HybridRNNs offer a powerful framework for modeling behavior that is both predictable and interpretable, bridging the gap between cognitive models and neural network approaches.

1 Introduction

The ability to learn flexibly is crucial in a dynamic world [MacDowell et al., 2022]. In cognitive neuroscience, reversal learning tasks are widely used to study behavioral flexibility [Groman et al., 2019, Rudebeck et al., 2013]. Unlike classical reward learning, where contingencies remain stable, reversal learning requires participants to adapt when reward contingencies switch. The speed with which behavior shifts to the newly rewarding option reflects an individual's flexibility in reward learning [Bartolo and Averbeck, 2020].

To explain such flexibility, researchers have developed computational models including reinforcement learning (RL) [Daw et al., 2011, Farashahi et al., 2017, Wilson and Collins, 2019, Eckstein et al., 2022] and Bayesian inference [Costa et al., 2015, Eckstein et al., 2022]. Here, we use "RL" to refer to trial-by-trial learning driven by reward prediction errors, including learning processes implemented by classical models such as Rescorla–Wagner [Rescorla, 1972]. These models provide interpretable accounts of learning with simple update rules and few parameters. Extensions such as adaptive learning rates or value resets improve flexibility [Hauser et al., 2014, Sidarus et al., 2019, Barnby et al., 2022], yet hand-crafted models often fail to capture behavioral richness [Peterson et al., 2021, Miller et al., 2024, Eckstein et al., 2024], and their predefined assumptions risk introducing bias [Feher da Silva and Hare, 2020].

An alternative is to model behavior directly with artificial neural networks (ANNs), which make fewer structural assumptions and can approximate a wide range of strategies. Prior work has shown ANNs can successfully capture human decision-making, reward learning, and cognitive flexibility [Dezfouli et al., 2019, Song et al., 2021, Jaffe et al., 2023]. However, their large number free parameters challenges interpretability, motivating new methods to combine predictive accuracy with mechanistic insights.

*Equal contribution

Preprint.

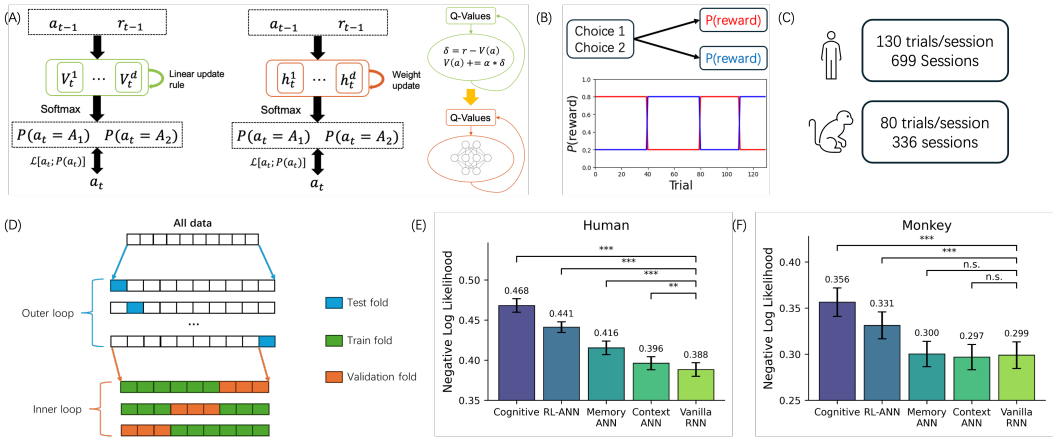

Figure 1: Task design and model performance. (A) Traditional cognitive modeling neural network modeling on cognition, and replacing updating rule with neural network. **(B)** Reversal learning task **(C)** Human and primate dataset overview. **(D)** Overview of inner-outer training loop. **(E and F)** Model comparison results (NLL, lower is better). Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, n.s. $p \geq 0.05$.

Recent approaches such as DisRNN [Miller et al., 2024], Tiny RNN [Ji-An et al., 2025], and HybridRNN [Eckstein et al., 2024] address this trade-off. We adopt the HybridRNN framework, which replaces hand-crafted RL update rules with learnable neural components while retaining interpretable structure. This allows integration of additional cognitive factors—such as context [Palminteri et al., 2015] and memory processes beyond RL [Collins and Frank, 2012, Davidow et al., 2016, Gershman and Daw, 2017]—to assess their contributions to flexible learning.

In this study, we fitted HybridRNN models to human and primate reversal learning behavior. These models allowed us to bridge predictive accuracy with interpretability. Among them, the Context-ANN provided the best account of behavior, capturing trial-by-trial adaptation in reversals. Analyses of hidden dynamics further revealed structured activity patterns consistent with flexible value updating. Together, our work positions HybridRNNs as a framework for linking model fit with insights into the computations underlying adaptive learning.

2 Main Findings

Models. We implemented five models spanning the range from classic to neural approaches. (i) *Simple RL*: a Rescorla–Wagner learner with two free parameters, learning rate α and inverse temperature β. (ii) *RL-ANN*: a hybrid model with the same structure as Simple RL, but in which the hand-crafted update rules (value, counterfactual, perseverance channels) are replaced by small MLPs that output additive updates to Q (and a perseverance bias), followed by a softmax for choice. (iii) *Context-ANN*: RL-ANN augmented with contextual inputs (the full value vector Q_{t-1} and previous perseverance c_{t-1}), enabling value normalization/context use [Palminteri et al., 2015]. (iv) *Memory-ANN*: RL-ANN augmented with the channels' previous hidden states to access a compressed history. (v) *Vanilla RNN*: a generic recurrent baseline that maps $[a_{t-1}, r_{t-1}]$ and s_{t-1} to action logits via a tanh RNN and softmax.

Training and evaluation. We used nested cross-validation to optimize negative log-likelihood (NLL): data for both humans and primates were split into 10 outer folds, and within each, a 3-fold inner CV selected hyperparameters. For each outer fold we refit on the nine training folds and evaluated NLL on the held-out fold, reporting the mean±s.e.m. across folds. Final models for dynamical analyses were retrained on the full dataset with the selected settings.

From RL to RNN: Context-ANN Nearly Closes the Gap We first modeled the human dataset using two extremes—Simple RL (Fig. 2B; parameters α, β) and a vanilla RNN. Predictive accuracy improved from Simple RL (mean NLL $= 0.468$) to the RNN (0.388; paired $t(9) = 19.65, p < 0.001$), indicating that classic RL misses systematic structure. Although Vanilla RNN achieved high predictive

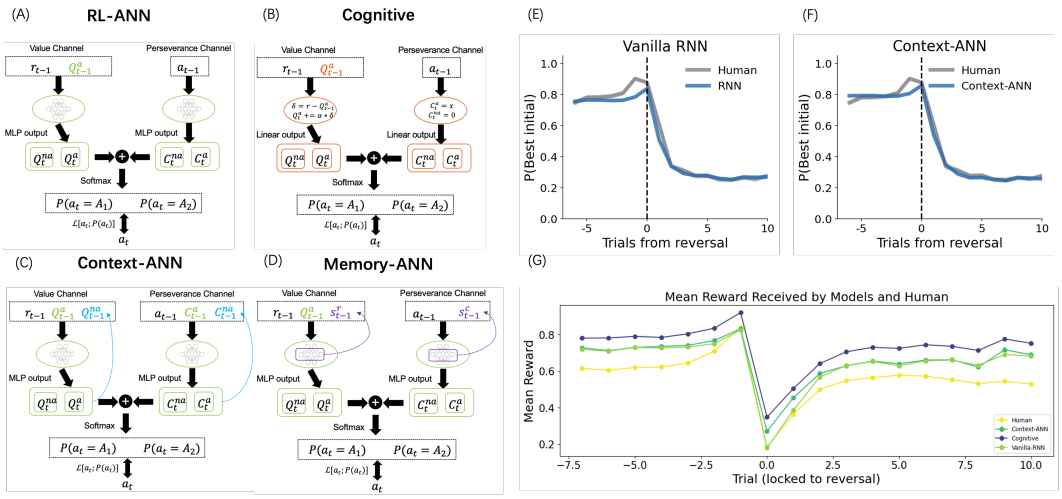

Figure 2: Task design and model performance. (A) - (D) Detailed model structure (see Appendix Methods for structure descriptions). **(E), (F)** Probability of staying at initial best choice for human, Vanilla-RNN and Context-ANN. **(G)** Mean reward received by models and human.

accuracy in modeling human behavior, its large number of fitted parameters makes interpretation challenging. To retain interpretability while improving fit, we progressively relaxed the constraints on the exact shape of the RL update. RL-ANN (Fig. 2A) improved over Simple RL (NLL $= 0.441$; $t(9) = 5.39$, $p < 0.001$), but still underperformed the RNN ($t(9) = -15.74$, $p < 0.001$). Adding counterfactual updating (Context-ANN, Fig. 2C) yielded a large gain (NLL $= 0.396$; vs. RL-ANN: $t(9) = -16.39$, $p < 0.001$), consistent with updating of non-chosen values [Palminteri et al., 2016, Eckstein et al., 2022], showing that human learners adjust the values of choices whose outcomes they have not directly observed. Adding choice perseverance (Memory-ANN, Fig. 2D) also helped (NLL $= 0.416$; $t(9) = -10.30$, $p < 0.001$). Among these, Context-ANN (NLL $= 0.396$) came closest to the RNN, but remained inferior ($t(9) = -4.52$, $p < 0.01$). Model selection favored Context-ANN with a perseverance channel (Fig. 1E). We also trained the same model family on the macaque dataset, using the identical nested cross-validation. Results mirrored the human data (Fig. 1F): the cognitive model performed worst (mean NLL $= 0.356$) and RL-ANN improved upon it (0.331, paired $t(9) = 10.74$, $p < 0.001$). Unlike the human results, Memory-ANN (0.300), Context-ANN (0.297), and vanilla RNN (0.299) were statistically indistinguishable from one another (paired t-tests; $df = 9$, all $p > 0.01$), and all three significantly outperformed RL-ANN (paired t-tests; $df = 9$, all $p < 0.0001$).

Context-ANN Captures Human Behavior Beyond quantitative model comparison, we examined whether the Context-ANN can reproduce humans' flexible learning in the reversal learning task. Figure. 2E, F shows the probability of repeating the initial choice (P_{initial}) aligned to reversal trials. Both the vanilla RNN and the Context-ANN capture the sharp drop in P_{initial} following contingency changes, closely matching human behavior. Furthermore, when aligning trials to the switch point, the Context-ANN also reproduces the human reward acquisition trajectory more faithfully than the cognitive baseline, indicating that the model captures not only overall choice patterns, but also trial-by-trial adaptation after reversals.

Context-Dependent Flexible Value Updating Strategy in Context-ANN To probe why the Context-ANN achieves superior predictive accuracy, we examined its internal value-updating mechanism. Specifically, we extracted the trained MLP responsible for value processing and simulated its responses across combinations of choices, rewards, and initial Q-values, yielding a vector-field representation (Fig. 3B). Unlike the linear cognitive model, whose attractors always lie at the extreme values of the rewarding option (Fig. 3A), the Context-ANN exhibits a qualitatively distinct structure: in the absence of reward (top row), attractors align along the diagonal, whereas upon receiving reward (bottow row), they shift toward the rewarding choice. This context-dependent attractor geometry highlights a unique computation not captured by classical models. Strikingly, our results parallel

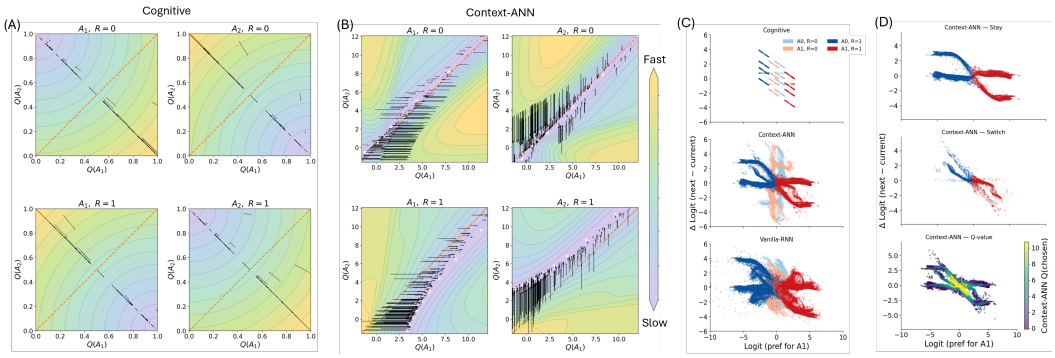

Figure 3: Dynamical-System Analyses Reveal the Advantages of Context-ANN. (A) Vector field of the cognitive model fitted to human data. Axes are the internal beliefs for the two choices. White crosses mark attractors; the orange dashed line shows $y{=}x$. The colorbar encodes the speed of Q-value change. Black arrows show one-trial belief updates predicted on the human dataset (proportionally subsampled across trials). **(B)** Same as (A), but for the Context-ANN. **(C)** Phase portraits of logit dynamics for the Cognitive model (top), Context-ANN (middle), and Vanilla-RNN (bottom). Colors indicate subjects' action and reward. **(D)** Decomposition of Context-ANN phase portraits on rewarded trials. Top: stay trials; middle: switch trials; bottom: Q-value heatmap.

recent findings by [Ji-An et al., 2025], who reported similar non-linear attractor dynamics when fitting human behavior with GRU networks, underscoring the biological plausibility of the Context-ANN's mechanism.

Phase portraits and Q-values reveal interpretable dynamics in Context-ANN Phase portraits of logit preference dynamics ($\Delta L = L_{t+1} - L_t$ vs. L_t) revealed systematic differences across models (Fig. 3C–D). The logit was defined as $L_t = \log \frac{p(A_0)}{p(A_1)}$, where $p(A_0)$ and $p(A_1)$ are the model's predicted choice probabilities. Context-ANN showed the clearest structure with horizontal attractor arms, slanted drifting arms, and small wings near indifference ($L_t \approx 0$, equal preference between A_0 and A_1; Fig. 3C, middle). Vanilla-RNN produced broader, less distinct arms (Fig. 3C, bottom), while the Cognitive model reduced to straight segments (Fig. 3C, top). Decomposing Context-ANN revealed that switch trials (action at next trial switches to another choice) concentrated on the small wings near indifference(Fig. 3D, middle), reflecting abrupt preference shifts when choices compete, while stay trials (trials maintaining the same action) gave rise to horizontal arms, representing stable attractors, and slanted arms, representing drifting stays where preference continues to shift despite repeating the same action, often near reversals or under inconsistent feedback (Fig. 3D, top). Overlaying Q-values further separated these regimes: high values appeared near indifference and drifting states, where uncertainty and competing choices make outcomes most informative, while low values marked horizontal ends, where preferences stabilize and updating ceases(Fig. 3D, bottom).

3 Conclusion and Limitation

Using HybridRNNs, we investigated the mechanisms of flexible learning in humans and primates. The Context-ANN revealed a distinctive strategy of value updating: its dynamics exhibited a context-dependent, non-linear attractor structure that shifts with reward feedback, closely paralleling trial-by-trial adaptation in reversal learning. These findings highlight a cognitive process underlying behavioral flexibility, showing how interpretable neural-cognitive models can uncover the computational principles of adaptive learning. However, reversal learning is a narrow task domain, and it remains unclear how well HybridRNNs generalize to more complex, naturalistic settings.

References

Camden J MacDowell, Sina Tafazoli, and Timothy J Buschman. A goldilocks theory of cognitive control: Balancing precision and efficiency with low-dimensional control states. *Current Opinion in Neurobiology*, 76: 102606, 2022.

Stephanie M Groman, Colby Keistler, Alex J Keip, Emma Hammarlund, Ralph J DiLeone, Christopher Pittenger, Daeyeol Lee, and Jane R Taylor. Orbitofrontal circuits control multiple reinforcement-learning processes. *Neuron*, 103(4):734–746, 2019.

Peter H Rudebeck, Richard C Saunders, Anna T Prescott, Lily S Chau, and Elisabeth A Murray. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nature neuroscience*, 16(8): 1140–1145, 2013.

Ramon Bartolo and Bruno B Averbeck. Prefrontal cortex predicts state switches during reversal learning. *Neuron*, 106(6):1044–1054, 2020.

Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

Shiva Farashahi, Christopher H Donahue, Peyman Khorsand, Hyojung Seo, Daeyeol Lee, and Alireza Soltani. Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron*, 94(2): 401–414, 2017.

Robert C Wilson and Anne GE Collins. Ten simple rules for the computational modeling of behavioral data. *Elife*, 8:e49547, 2019.

Maria K Eckstein, Sarah L Master, Ronald E Dahl, Linda Wilbrecht, and Anne GE Collins. Reinforcement learning and bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience*, 55:101106, 2022.

Vincent D Costa, Valery L Tran, Janita Turchi, and Bruno B Averbeck. Reversal learning and dopamine: a bayesian perspective. *Journal of Neuroscience*, 35(6):2407–2416, 2015.

Robert A Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current theory and research/Appleton-Century-Crofts*, 1972.

Tobias U Hauser, Reto Iannaccone, Juliane Ball, Christoph Mathys, Daniel Brandeis, Susanne Walitza, and Silvia Brem. Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA psychiatry*, 71(10):1165–1173, 2014.

Nura Sidarus, Stefano Palminteri, and Valérian Chambon. Cost-benefit trade-offs in decision-making and learning. *PLoS computational biology*, 15(9):e1007326, 2019.

Joseph M Barnby, Mitul A Mehta, and Michael Moutoussis. The computational relationship between reinforcement learning, social inference, and paranoia. *PLoS computational biology*, 18(7):e1010326, 2022.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372 (6547):1209–1214, 2021.

Kevin Miller, Maria Eckstein, Matt Botvinick, and Zeb Kurth-Nelson. Cognitive model discovery via disentangled rnns. *Advances in Neural Information Processing Systems*, 36, 2024.

Maria K Eckstein, Christopher Summerfield, Nathaniel Daw, and Kevin J Miller. Hybrid neural-cognitive models reveal how memory shapes human reward learning, 2024.

Carolina Feher da Silva and Todd A Hare. Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, 4(10):1053–1066, 2020.

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.

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.

Paul I Jaffe, Russell A Poldrack, Robert J Schafer, and Patrick G Bissett. Modelling human behaviour in cognitive tasks with latent dynamical systems. *Nature Human Behaviour*, 7(6):986–1000, 2023.

Li Ji-An, Marcus K Benna, and Marcelo G Mattar. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, pages 1–9, 2025.

Stefano Palminteri, Mehdi Khamassi, Mateus Joffily, and Giorgio Coricelli. Contextual modulation of value signals in reward and punishment learning. *Nature communications*, 6(1):8096, 2015.

Anne GE Collins and Michael J Frank. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7):1024–1035, 2012.

Juliet Y Davidow, Karin Foerde, Adriana Galván, and Daphna Shohamy. An upside to reward sensitivity: the hippocampus supports enhanced reinforcement learning in adolescence. *Neuron*, 92(1):93–99, 2016.

Samuel J Gershman and Nathaniel D Daw. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, 68(1):101–128, 2017.

Stefano Palminteri, Emma J Kilford, Giorgio Coricelli, and Sarah-Jayne Blakemore. The computational development of reinforcement learning during adolescence. *PLoS computational biology*, 12(6):e1004953, 2016.

Ioana Calangiu, Sepp Kollmorgen, John Reppas, and Valerio Mante. Prospective and retrospective representations of saccadic movements in primate prefrontal cortex. *Cell Reports*, 44(2):115289, 2025. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2025.115289. URL `https://www.sciencedirect.com/science/article/pii/S2211124725000609`.

Salman E Qasim, Aarushi Deswal, Ignacio Saez, and Xiaosi Gu. Positive affect modulates memory by regulating the influence of reward prediction errors. *Commun Psychol*, 2(1):52, June 2024.

Jessica V Schaaf, Laura Weidinger, Lucas Molleman, and Wouter van den Bos. Test-retest reliability of reinforcement learning parameters. *Behav Res Methods*, 56(5):4582–4599, September 2023.

A Methods

A.1 Tasks and datasets

Dataset 1: Adolescent Stochastic Reversal Task (Eckstein et al., 2022) This dataset comprises behavioral data from a large cross-sectional sample (ages 8–30, $n = 291$) performing a stochastic two-armed bandit reversal task [Eckstein et al., 2022]. On each trial, participants chose between two boxes, one of which yielded a probabilistic reward (75% vs. 0%). The rewarded side reversed unpredictably, requiring participants to balance persistence against flexibility. After preprocessing, we retained 305 valid sessions. The dataset is availible at https://osf.io/jm2c8/.

Dataset 2: Macaque Prefrontal Cortex Reversal Task (Bartolo & Averbeck, 2020) This dataset contains behavioral and high-density neural recordings from two macaques performing a reversal learning task [Bartolo and Averbeck, 2020]. Monkeys selected between two targets with unequal reward probabilities (0.7 vs. 0.3), which reversed stochastically within each block. Block structure varied between "What" (image-based contingencies) and "Where" (location-based contingencies) conditions. Neural population recordings (up to 1,000 simultaneously) from dorsolateral prefrontal cortex (dlPFC) enable analyses of Bayesian state inference and neural correlates of abrupt choice switching. After preprocessing, we retained 190 valid sessions.

Dataset 3: Saccadic Movement Representations (Calangiu et al., 2025) This dataset includes dlPFC neural recordings from three macaques across a variety of saccade-based tasks [Calangiu et al., 2025]. Animals performed visually guided, instructed saccade tasks with randomized delay periods, along with freely initiated saccades. Each trial required fixation, a rewarded saccade to a target, and subsequent fixations. The design allows separation of pre-saccadic (prospective) and post-saccadic (retrospective) neural representations, enabling study of how the dlPFC integrates memory and planning across contexts. After preprocessing, we retained 147 valid sessions. This dataset is released under the Creative Commons Attribution 4.0 International license and is available at https://zenodo.org/records/14360532.

Dataset 4: Reward Prediction Errors and Memory (Qasim et al., 2024) This human dataset ($n = 206$ participants) was collected online via a decision–memory paradigm [Qasim et al., 2024]. Participants performed a probabilistic two-armed bandit task (win probabilities 0.8 vs. 0.2, reversing

every 12 ± 1 trials). After each choice, participants viewed a unique face stimulus drawn from a perceptual memorability database. A subsequent recognition memory task tested recall of these faces against novel lures. The design disentangles contributions of reinforcement learning signals (reward prediction errors, RPEs) and perceptual memorability to memory encoding and retrieval, further modulated by affective state. After preprocessing, we retained 206 valid sessions. This dataset is released under MIT license and is availible at https://osf.io/awu3m/.

Dataset 5: Human Online Bandit and Reversal Learning (Schaaf et al., 2023) This dataset contains behavioral data from two independent online cohorts tested on reinforcement learning tasks [Schaaf et al., 2023]. One cohort ($n = 142$) performed a two-armed bandit task with probabilistic feedback under gain and loss blocks. Another cohort ($n = 154$) performed a reversal learning task where reward contingencies reversed across blocks. Both tasks were incentivized and repeated across two sessions (five-week interval), allowing assessment of test–retest reliability of reinforcement learning parameters. After preprocessing, we retained 94 valid sessions. The dataset is availible at https://osf.io/pe23t/.

Final training data. For model training, we grouped datasets by species and standardized trial lengths using padding. The human datasets (Eckstein, Qasim, Schaaf) yielded a total of 605 valid sessions, which were segmented into 699 sessions of 130 trials each. The monkey datasets (Bartolo & Averbeck; Calangiu) yielded a total of 337 valid sessions, which were segmented into 337 sessions of 80 trials each. These padded sessions constitute the final datasets used for training and evaluation of our models.

A.2 Model Architecture

Reinforcement Learning Models To identify the best classic cognitive model, we systematically compared reinforcement learning models with different structures [Wilson and Collins, 2019]. The Rescorla-Wagner (RW) model assumes that learning is driven by unexpected outcomes, including surprising occurrence or omission of reward during associative learning [Rescorla, 1972]. In a RW model, the prediction error (PE) signal defines the difference between observed and expected reward. This PE signal is weighted by a learning rate parameter when individuals update their reward expectations. According to this model, the agent's expected reward for the chosen action on trial t, $Q(t)$ is calculated as follows:

$$Q_t(a) = Q_{t-1}(a) + \alpha[r_{t-1} - Q_{t-1}(a)] \tag{1}$$

In this equation, $Q_t(a)$ refers to the reward value expectation for the chosen option at trial t, Q_{t-1} is the reward value expectation for the chosen option at trial $t-1$, the r_{t-1} (given that rewards were normalized to 0 and 1) refers to the reward actually perceived by participants at trial $t-1$, and the α represents the learning rate. We call this processing step the

value channel because it processes reward-related value information.

At trial t, for action selection, the vector $Q(t)$ of all action values is transformed into a vector of choice probabilities p_t of the same length using the softmax function. The transformation is determined by a free parameter "inverse decision temperature" β, to induce more deterministic or more random choices:

$$p_t(a) = \frac{e^{\beta Q_t(a)}}{\sum_{a'} e^{\beta Q_t(a')}} \tag{2}$$

For increasingly flexible RL models, we followed the work of Eckstein et al. [2022] to include counterfactual value updating and choice persistence into the framework. With counterfactual value updating, in each trial t, the value of the non-chosen action is also updated, using an "imaginary" counterfactual reward that is the inverse of the actually received reward $1 - r_{t-1}$ (given that rewards were normalized to 0 and 1), and based on the same learning rate as the chosen action:

$$Q_t(na) = Q_{t-1}(na) + \alpha[(1 - r_{t-1}) - Q_{t-1}(na)] \tag{3}$$

In this equation, $Q_t(na)$ refers to the reward value expectation for the non-chosen option at trial t. We call this additional, parallel processing step the "counterfactual value channel".

Finally, we extended the RL models with

perseverance, which enables action repetition independently of rewards. The

perseverance term c adds a small

bonus (of size κ) to the value of the action a that was chosen on the previous time step, but not to all other actions na:

$$\begin{aligned} c_t(a) &= \kappa \\ c_t(na) &= 0 \end{aligned} \tag{4}$$

RL-ANN Model The general idea of creating RL-ANN is to replace the value and perseverance updating function from Best RL with neural networks. For the three

channels designed in classic RL models, their hand-crafted updating functions (see Equations (1,2,3)) were replaced with a simple neural network.

For each trial, RL-ANN's value channel takes the reward r_{t-1} and the value of chosen action Q_{t-1} from the previous trial as inputs, just like the Simple RL model:

$$i_t^{(r)} = [Q_{t-1}(a), r_{t-1}]$$

The hidden layer activations, referred to as the "state" $s_t^{(r)}$, are computed by processing the input vector through the network's first fully-connected layer. Specifically, the input vector is multiplied by the weight matrix $W_1^{(r)}$, the bias term $b_1^{(r)}$ is added, and the result is passed through a tanh activation function to introduce non-linearity:

$$s_t^{(r)} = \tanh\left(W_1^{(r)} i_t^{(r)} + b_1^{(r)}\right)$$

The output of value channel is the update to the value of chosen action $Q_t(a)$, and is calculated by passing the state $s_t^{(r)}$ through the second fully-connected layer with weights $W_2^{(r)}$ and biases $b_2^{(r)}$:

$$\begin{aligned} u_t(a) &= W_2^{(r)} s_t^{(r)} + b_2^{(r)} \\ Q_t(a) &= Q_{t-1}(a) + u_t(a) \end{aligned}$$

The only difference between Simple RL and RL-ANN's value channels hence is that we replaced the linear updating function from Simple RL to neural networks in RL-ANN.

We next turned to the counterfactual value channel, which has the same structure as the value channel: it takes previous reward r_{t-1} and value of non-chosen action $Q_{t-1}(na)$ as input and predicts the update to the current value of the non-chosen action:

$$\begin{aligned} i_t^{(na)} &= [Q_{t-1}(na), r_{t-1}] \\ s_t^{(na)} &= \tanh\left(W_1^{(na)} i_t^{(na)} + b_1^{(na)}\right) \\ u_t(na) &= W_2^{(r)} s_t^{(r)} + b_2^{(r)} \\ Q_t(na) &= Q_{t-1}(na) + u_t(na) \end{aligned}$$

Hence, RL-ANN's counterfactual channel has the same structure as Simple RL's.

The perseverance channel is also a 3-layer fully-connected multilayer perceptron (MLP). The input to the channel is the action from the previous trial a_{t-1}, and the output of the channel is a vector $c_t \in \mathbb{R}^2$. For binary choice tasks, each dimension of c_t represents the perseverance scalar corresponding to one of the two available actions.

$$\begin{aligned} i_t^{(a)} &= a_{t-1} \\ s_t^{(a)} &= \tanh\left(W_1^{(a)} i_t^{(a)} + b_1^{(a)}\right) \\ c_t &= W_2^{(a)} s_t^{(a)} + b_2^{(a)} \end{aligned}$$

Finally, depending on the need to combine the channels, the values $Q_t(a)$, counterfactual values $Q_t(na)$, and perseverance C_t are combined additively and then pass through the softmax to select an action for the next trial:

$$\begin{aligned} h_t &= Q_t(a) + Q_t(na) + c_t \\ p_t &= \text{softmax}(h_t) \end{aligned}$$

Context-ANN Model Context-ANN extends RL-ANN by adding additional context information as input to the network, and is the winning model in the current paper. We provide context information for value channel with the vector Q_{t-1} and perseverance channel with the vector c_{t-1} as additional inputs, giving rise to access of value and perseverance of all possible choices [Palminteri et al., 2015].

$$i_t^{(r)} = [Q_{t-1}(a), r_{t-1}, Q_{t-1}]$$

$$i_t^{(a)} = [a_{t-1}, c_{t-1}]$$

Besides the difference in model inputs, the model structure and outputs stay the same with RL-ANN.

Memory-ANN Model In another approach of extending RL-ANN, we added states of hidden units from previous trial $s_{t-1}^{(r)}$ and $s_{t-1}^{(a)}$ as additional inputs. This modification of input enables the model to access its own compressed representation of the past history.

$$i_t^{(r)} = \left[Q_{t-1}(a), r_{t-1}, s_{t-1}^{(r)}\right]$$

$$i_t^{(a)} = \left[a_{t-1}, s_{t-1}^{(c)}\right]$$

Vanilla RNN Model Vanilla RNN is a simple, fully-connected recurrent neural network with three layers. On each trial t, it receives the previous action a_{t-1} (one-hot encoded) and reward r_{t-1} as input. The hidden state s_t is updated based on the current input and the previous hidden state. This hidden state is then passed through an output layer to generate action logits, which are converted into probabilities using a softmax function:

$$i_t = [a_{t-1}, r_{t-1}]$$

$$s_t = \tanh(W_1 i_t + b_1 + W_r s_{t-1})$$

$$h_t = W_2 s_t + b_2$$

$$p_t = \text{softmax}(h_t)$$

A.3 Model Training

Training and evaluation We used nested cross-validation with negative log-likelihood (NLL) as both the selection criterion and the evaluation metric,

$$L = -\sum_{i=1}^{bs} \sum_{t=1}^{n_{\text{trials}}} \log(p(a_{t,i}))$$

where bs is the batch size and $n_{\text{trials}} = trials$.

The data were shuffled and then split into ten folds. In each outer fold, one fold (10%) served as a test set, and the remaining nine folds (90%) formed the training pool. Within this pool, we performed a 3-fold inner cross-validation: split 90% into three parts, train on two and validate on one, rotate and average the validation NLL. Hyperparameters with the lowest average validation NLL were chosen for that outer fold. We then retrained on all nine folds and evaluated NLL on the held-out test fold. Repeating over all ten folds yielded ten test NLLs; we reported their mean and standard error of the mean.

We considered hidden sizes of 8, 16, 32, or 64 units; learning rates of 0.001 or 0.0001; weight decay of 0.001 or 0.0001; and batch sizes of 32 or 64. We used Adam optimizer.

We selected the configuration with the best average test NLL across the ten outer folds and retrained on 100% of the data. This final model was used for the analysis of internal dynamics.

Compute resources We ran our code on Google Colab using a v6e-1 TPU. Training a single model once required approximately 1–2 minutes. For the complete inner–outer loop over the described parameter grid, training one model took about 12 hours.

A.4 Qualitative Model Comparison

Reversal Time We aligned trials to each reversal ($t = 0$), extracting a window of seven pre- and ten post-reversal trials. On this aligned axis, we computed, for humans and each model, (i) the probability of repeating the initial choice (the correct choice before reversal) and (ii) the mean reward obtained.

A.5 Dynamics Analysis

Vector field analysis of inner belief We analyzed the dynamics of the internal belief for the two actions in the cognitive model and the Context-ANN by visualizing vector fields of belief updates. Vanilla RNN, Memory-ANN, and RL-ANN lack a direct belief representation and were therefore excluded. To examine condition-specific updates, we stratified trials by action and reward and plotted the corresponding belief-change vectors. For each plot, we set the belief axes to the empirical minimum and maximum and discretized this range into a uniform 40×40 grid to render the vector field.

Dynamics analysis of choice favor We quantify action preference by the log-odds of choosing A_1 over A_2:

$$\ell_t \;=\; \log\frac{p_t(A_1)}{p_t(A_2)}, \qquad \Delta\ell_t \;=\; \ell_t - \ell_{t-1}.$$

Trials are stratified by the previous action and outcome (a_{t-1}, r_{t-1}). We examine how $\Delta\ell_t$ depends on these conditions and on the prior preference ℓ_{t-1}. We also assess value influences by relating $\Delta\ell_t$ to the prior internal belief difference $\Delta Q_{t-1} = Q_{t-1}(A_1) - Q_{t-1}(A_2)$.

B Broader Impact

Our paper investigates possible learning mechanisms underlying human reversal learning. The potential social impact is to improve our understanding of human learning processes and contribute to a deeper knowledge of brain function. At present, we do not foresee any negative societal impacts.