

# Learning under noisy supervision is governed by a feedback–truth gap

Elan Schonfeld<sup>1,\*</sup> Elias Wisnia<sup>2</sup>

<sup>1</sup>Department of Biology, Columbia University, New York, NY, USA

<sup>2</sup>Department of Chemistry, Columbia University, New York, NY, USA

\*Corresponding author: elan.schonfeld@columbia.edu

## Abstract

When feedback is absorbed faster than task structure can be evaluated, the learner will favor feedback over truth. A two-timescale model shows this feedback–truth gap is inevitable whenever the two rates differ and vanishes only when they match. We test this prediction across neural networks trained with noisy labels (30 datasets, 2,700 runs), human probabilistic reversal learning ( $N = 292$ ), and human reward/punishment learning with concurrent EEG ( $N = 25$ ). In each system, truth is defined operationally: held-out labels, the objectively correct option, or the participant’s pre-feedback expectation – the only non-circular reference decodable from post-feedback EEG. The gap appeared universally but was regulated differently: dense networks accumulated it as memorization; sparse-residual scaffolding suppressed it; humans generated transient over-commitment that was actively recovered. Neural over-commitment ( $\sim 0.04\text{--}0.10$ ) was amplified tenfold into behavioral commitment ( $d = 3.3\text{--}3.9$ ). The gap is a fundamental constraint on learning under noisy supervision; its consequences depend on the regulation each system employs.

## Introduction

In many domains, learners do not receive truth as a direct instructional cue. Rather, learners update their internal representations of the world based upon feedback that only imperfectly reflects the truth. In machine learning, noisy labels can lead to memorization and poor generalizability<sup>1–4</sup>. In human decision-making, probabilistic outcomes can blur the distinction between contingency and chance, requiring the learner to act on noisy feedback<sup>5,6</sup>. While these settings seem quite different, both represent examples of noisy supervision. Several groups have argued for deeper integration between AI and neuroscience<sup>7–10</sup>; however, we have lacked a common quantitative framework for understanding how noisy supervision affects learning processes in artificial and biological systems.

Here, we describe such a framework – the feedback–truth gap – the extent to which a learner’s internal representation of the environment tracks feedback more than it tracks truth. We operationally define this term in three systems: for a neural network, the training-minus-validation accuracy difference as a function of epoch; for a human reversal task, the distance between feedback-aligned and truth-aligned responses to contingency reversals; and for an EEG experiment, the difference between post-feedback neural decoding of the feedback value and the participant’s pre-feedback expectation (see Methods). Critically, in the EEG system the truth reference is the participant’s prior expectation of correctness – an operational proxy, not the latent task state – because no non-circular reconstruction of objective environmental truth was decodable from the post-feedback signal (Extended Data Fig. 10). While the definitions vary across systems, they share the same

formal properties: the learner revises its internal state based upon feedback before it becomes possible to resolve the discrepancy between the learner’s representation and the true state of the environment. Memorization dynamics in neural networks have been studied extensively<sup>11;12</sup>, and so have perseverative errors after contingency reversals in animals and humans<sup>13–16</sup>. Computational models of human learning have focused on asymmetric learning rates for reward and punishment<sup>17–19</sup>. However, none of these previous studies have provided a necessary condition for when a learner will inevitably become over-committed to feedback, nor have they linked post-reversal over-commitment in humans to the generalization gap in machine learning.

Using a minimal two-timescale learner, we establish that a feedback–truth gap is inevitable when feedback is integrated on a faster timescale than the learner’s evaluation of the task structure, and that the gap disappears only when the two timescales are equal (Methods; Fig. 2). We then empirically test this prediction across three independent systems – neural networks trained with label noise (30 datasets, 2,700 runs), human probabilistic reversal learning ( $N = 292$ ), and human reward/punishment learning with concurrent EEG ( $N = 25$ ) – using the same gap metrics ( $\text{AUG}_{\text{pos}}$ ,  $T^*$ ) across all three systems.

We found that the gap appeared in all three systems but was regulated in very different ways (Fig. 1). In dense neural networks, the gap grew progressively and became a permanent feature of the network’s behavior – memorization. In sparse-residual architectures and noise-robust training, the gap was reduced or suppressed (Fig. 4). In human reversal learning, the gap increased immediately following contingency reversals and was subsequently decreased through active behavioral regulation; the relationship between the size of the gap and the recovery time varied among individuals (Fig. 3). In EEG, post-feedback activity represented feedback valence and prior expectations, and individual differences in the neural gap predicted the amount of behavioral commitment each individual exhibited (Fig. 5). The same measurable quantity that indexed damage to learning in unregulated neural networks had the opposite implications for performance in regulated human learning.

## Results

### The feedback–truth gap appears across learning systems

We quantified the feedback–truth gap in three independent systems: artificial neural networks, human probabilistic reversal learners, and human reward/punishment learners with concurrent EEG (Fig. 1). The gap represents the same formal property of the learner’s behavior – feedback-aligned minus truth-aligned performance – but was measured using each system’s own notions of feedback and truth. The gap is therefore a common measurement construct, not a claim of common mechanism.

In neural networks trained on tabular data with 40% symmetric label noise, training accuracy rose faster than validation accuracy, resulting in a persistent positive gap that grew throughout training (Fig. 1a). This gap resulted from progressive memorization of the noisy labels; the network tracked feedback – including noise – at the cost of generalizing to new, unseen data. No mechanism in either the training process or the network architecture prevented the growth of the gap. Across five benchmark datasets (Ionosphere, Glass, Sonar, WDBC, Vehicle) and six architectures, including label smoothing,  $L_2$  regularization, and residual variants as controls, the gap was observed in greater than 90% of model-noise combinations at 40% label noise. The gap increased with the label noise rate and disappeared completely when the label noise rate was 0% (Extended Data Figs. 1 and 2). The unregulated dense neural network thus provides the reference prediction of the two-timescale model, while the control conditions verify that the gap persists under conventional remediation and is not specific to any particular architecture.

In human probabilistic reversal learning ( $N = 292$ , ages 8–30; ref. 11), participants chose between

two options with 75%/25% reward probabilities that reversed periodically (empirical noise rate: 17.1%). We calculated the same gap metric – feedback-aligned minus truth-aligned choice accuracy – on a rolling trial-by-trial basis, synchronized with the contingency reversals. Following each reversal, feedback-aligned accuracy recovered more quickly than truth-aligned accuracy, resulting in a transient positive gap that peaked within approximately five trials and resolved over 15–20 trials (Fig. 1b). This signal was extremely reliable ( $\text{AUG}_{\text{pos}} = 0.049 \pm 0.032$ ,  $t(291) = 26.4$ ,  $d = 1.55$ ,  $p < 10^{-75}$ ) and nearly universal: all 292 subjects demonstrated positive over-commitment following a reversal, and 99.3% eventually regained stable truth-aligned responding (mean  $T^* = 9.9$  trials). Unlike in networks, the human gap resolves: active behavioral regulation closes it.

In concurrent EEG recordings during reward and punishment learning ( $N = 25$ ; ref. 12), we decoded post-feedback feedback valence and each subject's pre-feedback expectation from the post-feedback EEG (Methods). Only one of four candidate truth references – the participant's prior expectation of correctness – was reliably decodable in this window; objective reconstructions of the task state based on Bayesian and reinforcement-learning models were not decodable (Extended Data Fig. 10). The expectation-truth decoder performed above chance in both tasks (REW: AUROC = 0.541,  $t(23) = 2.43$ ,  $p = 0.012$ ; PUN: AUROC = 0.545,  $t(24) = 3.84$ ,  $p < 10^{-3}$ ). The primary endpoint is not individual decoder accuracy but the trial-by-trial gap between the two decoders – how much more confidently the EEG signal represents feedback valence than expected accuracy. The neural gap was positive in all subjects included in the analysis (reward:  $N = 24$ ; punishment:  $N = 25$ ; Extended Data Fig. 3), and feedback-dominant signals in the EEG diminished over trials while expectation-related signals followed feedback (Fig. 1c), consistent with progressive adaptation of the neural response.

Fig. 1 thus illustrates a hierarchy of regulation. Over-commitment to feedback occurred in all three systems, but the progression of the gap through time differed. Dense networks drifted further from truth; human choices returned to truth; and EEG signals adapted at an intermediate timescale. Two questions follow: why does the gap occur at all, and what determines whether the system regulates it?

### The gap is inevitable once timescales differ

We analyzed the minimal two-timescale learner to determine why the gap occurs so uniformly in our experiments. The model uses a fast feedback-integrating channel (rate  $\alpha_{\text{fast}}$ ) and a slower truth-integrating channel (rate  $\alpha_{\text{slow}}$ ) (Fig. 2; Methods). Noise in the feedback causes any change in the underlying state to expose the difference in timescales: the fast channel moves before the slow channel, producing a temporary separation of feedback tracking and truth tracking.

Using a closed-form solution (Supplementary Derivation), the gap after a state change of size  $\Delta$  at  $t = 0$  is:

$$\text{gap}(t) = \Delta \cdot (e^{-\alpha_{\text{slow}} \cdot t} - e^{-\alpha_{\text{fast}} \cdot t})$$

which is positive for all  $t > 0$  whenever the two rates differ (Fig. 2d). The peak gap magnitude depends on the timescale ratio  $r = \alpha_{\text{fast}}/\alpha_{\text{slow}}$ : it approaches  $\Delta$  as  $r$  grows large and shrinks to zero as  $r$  approaches 1. If the two timescales match ( $r = 1$ ), the gap is identically zero regardless of the noise level. In simulation, at  $r = 1$  with 20% noise, the mean peak gap stayed below  $10^{-5}$  (Fig. 2d); across 40 timescale ratios the correspondence between theory and simulation gave  $R^2 > 0.99$ .

Based on these findings, we created a phase diagram of the two-timescale learner with respect to timescale ratio and noise level (Extended Data Fig. 2c). The only no-gap boundary in timescale space is when the two channels operate at the same rate ( $r = 1$ ); simulations confirm the predicted gap sizes throughout the entire parameter range. All three empirical systems can be located on this diagram: dense networks fall in the high-ratio, high-noise region ( $r \approx 10\text{--}20$ ), explaining their

large persistent gaps; humans occupy a moderate-ratio regime ( $r \approx 3\text{--}5$ ) where regulation can close the gap; sparse-residual networks have reduced effective ratios and correspondingly smaller gaps. All systems examined empirically demonstrated that removing label noise collapsed the gap and eliminated the advantage of sparse bottlenecks (Extended Data Fig. 5).

This result describes what must occur based on the two-timescale structure; it does not describe how any individual system implements learning. Once the feedback is noisy and is integrated faster than truth, the only way to prevent transient over-commitment is to delay, suppress, or compensate the gap through downstream regulation. Because the gap is deterministic given the timescale ratio, it is predictable from early data: the first 15 epochs of network training suffice to classify the eventual memorization regime (epochs 35+) with AUC = 0.96, and early-to-late prediction also works in both human paradigms (Extended Data Fig. 4).

### The significance of the gap depends upon regulation

Although the same gap construct appeared in every system examined, the relationship of the gap to performance differed qualitatively among the systems (Fig. 3). The significance of the gap is determined entirely by whether and how the system regulates it.

In neural networks under synthetic label noise, the cumulative gap magnitude ( $\text{AUG}_{\text{norm}}$ ) was a strong negative predictor of validation accuracy across 150 dataset–model pairs ( $\beta = -0.641$ ,  $p < 10^{-6}$ , fixed-effects regression; Fig. 3a). Higher memorization resulted in lower generalization; the gap was detrimental. Under natural noise (crowd-labelled CIFAR-10N), this relationship weakened as model rankings compressed, and onset timing  $T^*$  became the informative metric ( $r = 0.672$ ,  $p = 0.048$ ; Extended Data Fig. 2), demonstrating regime dependence in which metrics carry diagnostic information.

The relationship was reversed in human reversal learning. Subjects who exhibited larger transient gaps after contingency reversals also exhibited faster subsequent recovery of truth-aligned responding (Fig. 3b). The gap did not indicate damage; instead, it indicated the extent to which the subject engaged with the changed contingency, and a larger initial discrepancy between feedback tracking and truth tracking was associated with more rapid realignment. This does not imply that the gap was beneficial; rather, it implies that the downstream effects of the gap depended on whether regulatory mechanisms existed to close it. This reversal (negative gap–performance relationship in networks, positive gap–recovery relationship in humans) is the central dissociation: the same measurable quantity had opposite implications for performance depending on regulatory capacity.

The EEG data identified a neural basis for this dissociation. The neural gap was spatially organized, with feedback-dominant components at frontal electrodes and expectation-dominant components at posterior sites (Extended Data Fig. 3). Instead of a uniform scalar measure, the post-feedback signal separated evaluative and expectation-related components (Fig. 3c), consistent with the established role of medial frontal cortex in feedback monitoring<sup>20,21</sup> and with the theta, P300, and ERN literatures on outcome evaluation<sup>22–24</sup>.

The results thus defined three regulatory regimes. Dense networks lacked regulation, and the gap accumulated without restriction, directly indexing damage. Additional noise-robust methods (co-teaching, forward loss correction) suppressed but did not abolish the gap (Extended Data Table 1). Architectural constraints limited the ability to overfit the feedback, thereby suppressing the gap. Humans exhibited transient gaps and then recovered through behavioral adjustment. Although the phenomenon occurred in every system, the nature of the regulatory response determined the outcome.

## Regulation is tunable and dissociates pressure from recovery

Two complementary analyses assessed whether the regulation of the gap could be systematically controlled: architectural intervention in neural networks and cross-level integration in human learners (Figs. 4–5).

The strength of a sparse-residual architectural bottleneck was varied ( $\alpha \in \{0.1, 0.25, 0.5, 1.0\}$ ) across five datasets with 40% label noise (10 seeds per configuration). Both  $\text{AUG}_{\text{norm}}$  and  $T^*$  increased monotonically with  $\alpha$  (Spearman  $\rho = 1.0$  for both metrics; Fig. 4a–b), demonstrating graded causal control over the memorization dynamics. At the dataset level, this monotonic suppression held in all datasets with measurable gaps ( $\text{AUG}_{\text{norm}} > 0.005$ ;  $\rho = 1.0$  in 3/3), whereas datasets at the noise floor demonstrated no systematic trends because there was no gap to suppress (Extended Data Fig. 5). However, test accuracy did not follow this monotonic pattern: it peaked at  $\alpha = 0.1$ – $0.25$  and decreased at higher values (Fig. 4c). This non-monotonic dissociation demonstrated a capacity–memorization trade-off in which configurations that maximized suppression of the gap did not maximize generalization. (For comparisons with label smoothing, weight decay, and loss-correction methods<sup>25,26</sup>, see Methods and Supplementary Information.) Additional validation across 30 OpenML datasets confirmed that sparse bottlenecks improved accuracy on 93% of datasets at 40% noise but provided no reliable advantage at 0% noise (Extended Data Fig. 5; Extended Data Table 2). Suppression is noise-dependent. Sparse-residual architectures served as a causal probe for graded, monotonic control over the gap, not as a proposed method for practical noise robustness.

Behavioral over-commitment to feedback was large and consistent in the EEG dataset ( $N = 25$ ; ref. 12). All 25 subjects in both reward and punishment conditions showed positive over-commitment: reward  $t(24) = 16.6$ ,  $d = 3.32$ ,  $p < 10^{-14}$ ; punishment  $t(24) = 19.6$ ,  $d = 3.92$ ,  $p < 10^{-15}$  (Fig. 5a). Decomposition into win-stay ( $\sim 0.80$ – $0.83$ ) and lose-shift ( $\sim 0.20$ – $0.23$ ) components demonstrated that over-commitment was largely due to high persistence in rewarding actions, not failure to switch after negative outcomes. The concurrent neural gap, as assessed by EEG decoders, was considerably smaller (neural  $\text{AUG}_{\text{pos}} \approx 0.04$ – $0.10$ ; Fig. 5b), demonstrating approximately tenfold amplification from neural representation to behavioral strategy.

The neural gap was highly reproducible (split-half correlation between even and odd trials:  $r = 0.35$ ,  $p < 10^{-5}$  for reward;  $r = 0.45$ ,  $p < 10^{-8}$  for punishment) and decreased from early to late trials in the punishment condition ( $d = 0.82$ ,  $p < 0.001$ ; Fig. 3d), suggesting that the neural gap tracks a learning-related process rather than noise or drift (Extended Data Fig. 3). At the individual level, subjects with larger neural gaps also exhibited stronger behavioral commitment (reward:  $\rho = 0.43$ ,  $p = 0.036$ ; punishment:  $\rho = 0.42$ ,  $p = 0.042$ ; Fig. 5c). Neither EEG channel significantly predicts single-trial behavioral choice; rather, both channels appear to provide noisy measures of the same slow-changing, task-dependent learning state. The amplification is concordant across levels and across individuals (see Extended Data Fig. 6 for further individual-differences analyses, Extended Data Fig. 9 for robustness to the class-balance exclusion threshold, and Extended Data Table 4 for reinforcement-learning parameter associations).

The gap is suppressed differently in each system. Architectural bottlenecks suppress the gap by reducing the capacity of the network. Human brains employ a dynamic process that amplifies modest neural over-representation into large behavioral commitments and then corrects them. While the gap exists in all three systems, the systems differ in how they regulate it.

## Discussion

The feedback–truth gap is inherently mathematical: when a learning system integrates feedback before it can verify the true state of affairs, over-commitment to the feedback is inevitable. This gap

was consistently evident in all of the systems evaluated. More importantly, the critical finding is that systems differ categorically in how they respond. Networks without regulation generate the gap as damage; architectural constraints produce static suppression; and humans generate active recovery, with concordant amplification at the neural and behavioral levels. This taxonomy – unregulated, suppressed, recovered – provides a conceptual framework for evaluating how different systems manage a common computational constraint.

**Scope.** We do not contend that humans learn in the same way as artificial neural networks. The shared aspect of the research is a framework for measurement – the feedback–truth gap – applied to the same formal manipulation (perturbing the agreement between feedback and ground truth at each learning step) across substrates. The gap is an inherent consequence of timescale mismatch, not something we advocate as adaptive or optimal. The sparse-residual architectures serve as causal probes; the human neural evidence is correlational. In the machine-learning studies, the architectural manipulations serve as controlled probes rather than proposed general-purpose noise-resistant methods; topology ablations indicate that the sparse-residual scaffolding, not the graph structure itself, suppresses the gap (Extended Data Table 3).

**Shared framework, not shared mechanism.** Convergence across substrates represents shared measurement, not shared mechanism. The same quantitative metrics ( $\text{AUG}_{\text{pos}}$ ,  $T^*$ , commitment indices) capture gap dynamics in network training curves, human reversal learning, and reward/punishment strategies. The formal structure is identical across domains: noisy supervision, time-varying commitment, and eventual alignment with truth. However, the substrate-specific process that generates and regulates the gap is different in each case, and convergent measurement does not imply convergent implementation.

**Regulation as the differentiator.** What converges is the principle that raw gap pressure is modulated before the gap is expressed behaviorally. In networks, architectural regularization delays and suppresses the gap. In humans, neural over-representation ( $\rho = 0.43$  in reward, 0.42 in punishment; both  $p < 0.05$ ) is amplified roughly tenfold into behavioral over-commitment. This concordance holds at subject and block timescales; at the single-trial level, both neural and behavioral signals provide noisy estimates of a shared, slow-changing, task-dependent learning state. In networks the regulatory mechanisms are weight decay and capacity constraints. In humans, prefrontal control circuits are the natural candidate, and dopaminergic signaling – which modulates learning from prediction errors<sup>27;28</sup> and differentially shapes reward versus punishment learning<sup>29–31</sup> – is a plausible substrate for the amplification we observe.

**EEG dissociation.** EEG provided a distinct dissociation: post-feedback features decode feedback valence and the participant’s pre-feedback expectation, but do not reconstruct the latent task state (Extended Data Fig. 10). Post-feedback ERPs, especially the P300, are responsive to surprise (outcome minus expectation), resulting in a residue of prior belief rather than a direct representation of the environment. Thus, the neural feedback–truth gap represents the balance between outcome evaluation and expectation. By the time this balance is reached, the prior beliefs are already partially regulated; feedback and expectation representations are nearly equalized, so the much larger behavioral commitment is likely mediated by downstream control rather than simply scaled versions of the same neural signal.

**Regime dependence.** Neither the gap nor its diagnostic value is uniform across conditions. In networks,  $\text{AUG}_{\text{norm}}$  predicts accuracy under synthetic noise but loses discrimination under natural noise, where  $T^*$  becomes informative instead. In humans, the neural gap is modest ( $d \approx 1.0$ ) but is amplified manyfold at the behavioral level ( $d = 3.3\text{--}3.9$ ), with significant cross-modal correlations in both tasks ( $\rho = 0.43, 0.42$ ;  $p < 0.05$ ). The phase diagram (Extended Data Fig. 2c) captures this continuous variation: gap magnitude scales with noise rate and timescale ratio and vanishes at specific boundaries. No single metric is universally diagnostic; the informative metric depends on

the noise regime. The capacity–memorization tradeoff connects to the double-descent literature<sup>32</sup>, mathematical theories of learning dynamics<sup>33</sup>, and work on architecture and generalization<sup>34</sup>. Episodic memory<sup>35</sup> may also play a role, particularly where long delays separate feedback from truth verification.

This study has several limitations. Primarily, the human neural evidence is correlational. An exploratory pharmacological reanalysis is consistent with dopaminergic modulation of gap regulation: in a double-blind crossover ( $N = 27$ ;<sup>36</sup>), the D2/D3 agonist cabergoline flattened within-session gap regulation relative to placebo ( $t(26) = 2.31$ ,  $d = 0.44$ , one-sided  $p = 0.015$ ; Wilcoxon  $p = 0.016$ ; Extended Data Fig. 7), and cumulative gap level showed a directionally consistent reduction ( $d = -0.32$ , one-sided  $p = 0.053$ ). However, this result is preliminary; proper causal tests will require stimulation or lesion studies. A developmental analysis of the reversal-learning dataset ( $N = 292$ ; Extended Data Fig. 8) indicates a shift from more flexibility in children to more persistent use of feedback in adults, but future studies should assess these relationships in more diverse populations. The human samples are Western adults performing laboratory tasks from public repositories; generalizability to naturalistic or developmental settings is untested. The two-timescale model determines when the gap should appear, not how specific systems create or manage it; attention, working memory, and normalization dynamics all modulate the gap in ways that timescale separation alone does not capture. Expanding the framework to reinforcement-learning agents, developmental populations where prefrontal regulation is still maturing<sup>37–39</sup>, and clinical groups with abnormal feedback processing<sup>40–44</sup> will provide evidence for the framework’s broader utility.

The gap between feedback and truth should arise wherever there is noisy feedback and latent truth: scientific inference, social learning<sup>45</sup>, organizational adaptation, cultural transmission<sup>46</sup>. The structure echoes foundational ideas in associative learning<sup>47;48</sup>. Understanding what regulates the gap – across architectures, development, and pathology – may provide insight into developing both robust machine learning and human adaptive behaviors.

## Methods

### Two-timescale learner model

The simplest model exhibiting a feedback–truth gap tracks a binary environment via two exponential-moving-average channels with update rates  $\alpha_{\text{fast}}$  and  $\alpha_{\text{slow}}$  ( $\alpha_{\text{fast}} > \alpha_{\text{slow}} > 0$ ). The observed feedback is noisy: on each trial, the observed signal matches the true state with probability  $1 - \varepsilon$ . After a state reversal of magnitude  $\Delta$  at  $t = 0$ , the gap between the fast and slow estimates is  $\text{gap}(t) = \Delta \cdot (e^{-\alpha_{\text{slow}} \cdot t} - e^{-\alpha_{\text{fast}} \cdot t})$ , strictly positive for all  $t > 0$  (Supplementary Derivation), with peak magnitude  $\text{gap}_{\max} = \Delta \cdot r^{-1/(r-1)} \cdot (1 - r^{-1})$  where  $r = \alpha_{\text{fast}}/\alpha_{\text{slow}}$ . Simulations used 20 seeds per parameter setting with  $\alpha_{\text{slow}} = 0.02$ , noise  $\in \{0.1, 0.2, 0.4\}$ , and  $r$  ranging from 1 to 63. Phase diagrams were constructed from  $20 \times 20$  grids of noise rate  $\times$  timescale ratio.

### Machine learning experiments

Five tabular datasets from UCI/OpenML (Ionosphere, Glass, Sonar, WDBC, Vehicle) were selected because memorization is clearly observable within 100 epochs at 40% symmetric label noise. An extended benchmark of 30 OpenML datasets across three noise rates, three noise-type configurations, and ten seeds (2,700 runs total) confirmed generality. CIFAR-10 with synthetic noise and CIFAR-10N (human annotator noise) served as boundary cases.

Fully connected networks provided the dense baseline. The sparse-residual variant combines an identity shortcut with a fixed sparse bottleneck wired as an expander graph, reducing the dense layer’s

multiply-accumulate operations to roughly 22%. Bottleneck strength  $\alpha \in \{0.1, 0.25, 0.5, 1.0\}$  controls the capacity of the sparse branch. Three control architectures isolate specific ingredients: dense with label smoothing (Dense+LS), dense with a residual shortcut but no sparsity (Dense+Residual), and dense with strong  $L_2$  (Dense+StrongReg).

Two metrics quantify the gap.  $AUG_{norm}$  is the normalized area under the positive portion of the training-minus-validation accuracy curve;  $T^*$  is the first epoch at which the gap exceeds  $\tau = 0.05$  for three consecutive epochs. Both were computed per random seed and averaged. Statistical inference relied on fixed-effects regression across 150 dataset–model pairs, Spearman correlations for monotonicity, and 95% confidence intervals throughout. Each configuration had 10 seeds in the Full stage.

## Human probabilistic reversal learning

Trial-level behavioral data from 292 participants (ages 8–30) who performed a probabilistic reversal learning task<sup>49</sup> were obtained from a publicly accessible repository (OSF; <https://osf.io/7wuh4/>). Participants played a two-armed bandit with 75%/25% reward probabilities that reversed periodically (4–9 reversals per session, 125–131 trials total). The empirical noise rate was 17.1%.

Rolling accuracy was computed over an 8-trial window independently for truth-aligned and feedback-aligned responses; the gap is the difference (feedback minus truth), time-locked to each reversal.  $AUG_{pos}$  is the baseline-corrected area under the positive portion of this curve from trials 0 to 25 post-reversal.  $T^*$  is the first post-reversal trial at which truth accuracy exceeded 65% for 15 consecutive trials.  $AUG_{pos}$  was tested against zero with a one-sample  $t$ -test, reporting Cohen’s  $d$  and bootstrap 95% CIs.

## Human reward/punishment learning with EEG

Behavioral and 32-channel EEG data from 26 healthy adults who performed separate reward and punishment probabilistic learning tasks<sup>50</sup> were obtained from OpenNeuro (ds004295). One subject was dropped owing to an EEG recording malfunction, leaving  $N = 25$  for behavioral analyses; reinforcement-learning model fits used  $N = 23$ . Artifact-based epoch rejection reduced the neural decoder samples to  $N_{REW} = 24$  and  $N_{PUN} = 25$ ; a stricter class-balance exclusion for the prevalence analysis in Fig. 2c yielded  $N = 22$  (see Extended Data Fig. 9 for sensitivity to this threshold).

Both tasks were two-armed bandits with ~65–70% contingency validity. Correct responses in the reward condition earned +10 cents (otherwise 0); incorrect responses in the punishment condition triggered an aversive noise burst. Each task ran for approximately 280 trials. Behavioral commitment was defined as win-stay minus lose-shift probability and tested against zero with a one-sample  $t$ -test.

EEG preprocessing consisted of bandpass filtering at 1–40 Hz, average referencing, epoching –200 to 600 ms surrounding feedback onset, baseline correction, and rejection of epochs exceeding  $\pm 150 \mu V$  (implemented in MNE-Python). Three feedback-locked features entered the decoder: theta power (4–8 Hz at FCz, 200–400 ms), frontal beta power (13–30 Hz, frontal channels, 200–400 ms), and P300 amplitude (Pz, 250–450 ms). Logistic regression decoders (3-fold cross-validation) predicted feedback valence and pre-feedback expectation separately; the neural gap on held-out folds was  $P(\text{feedback}+|\text{EEG}) - P(\text{expectation}+|\text{EEG})$ . Four candidate truth definitions were evaluated – slow exponential moving average of feedback, pre-feedback expectation ratings, Bayesian HMM-inferred correct side, and fitted RL model Q-values – and the pre-feedback expectation was selected as the only non-circular truth label decodable above chance in both conditions (Extended Data Fig. 10). Cross-modal associations were assessed with Spearman correlation at the subject level (behavioral commitment vs. neural  $AUG_{pos}$ ) and Pearson correlation on smoothed trial-level time series (15-trial

moving average).

### Dopaminergic modulation (Extended Data Fig. 7)

Study 2 of the Cavanagh & Frank dataset<sup>36</sup> included  $N = 27$  healthy adults in a double-blind crossover design. Each participant completed the Probabilistic Selection Task under cabergoline (1.25 mg, a D2/D3 agonist) and placebo in separate sessions, with order counterbalanced. Three stimulus pairs offered graded noise regimes (20%, 30%, 40%). The primary measure was the gap regulation slope – the linear trend of the feedback-minus-truth gap time series (20-trial rolling window) over trials – quantifying how fast the gap resolves within a session. AUG<sub>pos</sub> (cumulative over-commitment) served as a secondary measure. Drug effects were examined with a one-sided paired *t*-test (directional hypothesis: a dopamine agonist should flatten the regulation slope and reduce over-commitment).

### Statistical reporting

All tests are two-tailed unless otherwise noted; the exception is the dopaminergic modulation analysis (Extended Data Fig. 7), where a directional hypothesis was specified before examining the data. Effect sizes are Cohen’s *d* throughout: group mean divided by pooled SD for between-group comparisons, mean difference divided by SD of differences for paired and one-sample tests. Correlations are Spearman  $\rho$ , each accompanied by 95% CIs. Normality was assessed with Shapiro–Wilk tests; when distributions deviated, non-parametric alternatives were run alongside the parametric tests (Wilcoxon signed-rank for the drug analysis, Mann–Whitney *U* for developmental comparisons).

No multiple-comparison correction was applied to the primary confirmatory tests ( $AUG_{pos} > 0$  in each system), because each is a single pre-specified hypothesis tested independently per system. Exploratory correlations (individual differences, cross-modal, developmental) carry exact *p*-values without correction and should be interpreted accordingly. The only covariate in any analysis is the empirical noise rate in the individual-differences partial correlation (Extended Data Fig. 6b). Sample sizes were determined by the publicly available datasets; no power analyses were conducted in advance. “Truth” is defined differently across systems: in the ML and PRL analyses, truth is objective (held-out labels and the objectively better option, respectively). In the EEG analysis, the truth reference is the participant’s pre-feedback expectation of correctness – the only non-circular, non-feedback label decodable above chance in both conditions (Extended Data Fig. 10). This is an operational decision and does not imply that the participant’s confidence is equivalent to objective truth.

### Data Availability

All datasets are publicly available: Eckstein *et al.* (2022) PRL data (<https://osf.io/7wuh4/>), Stoltz *et al.* (2022) EEG data (<https://openneuro.org/datasets/ds004295>), Cavanagh & Frank (2014) PST data (<https://openneuro.org/datasets/ds004532>), tabular datasets via UCI/OpenML, CIFAR-10N (<https://github.com/UCSC-REAL/cifar-10-100n>). Processed analysis artifacts are provided in the companion repository.

### Code Availability

Analysis scripts for all results are available at [GitHub repository URL upon acceptance].

## References

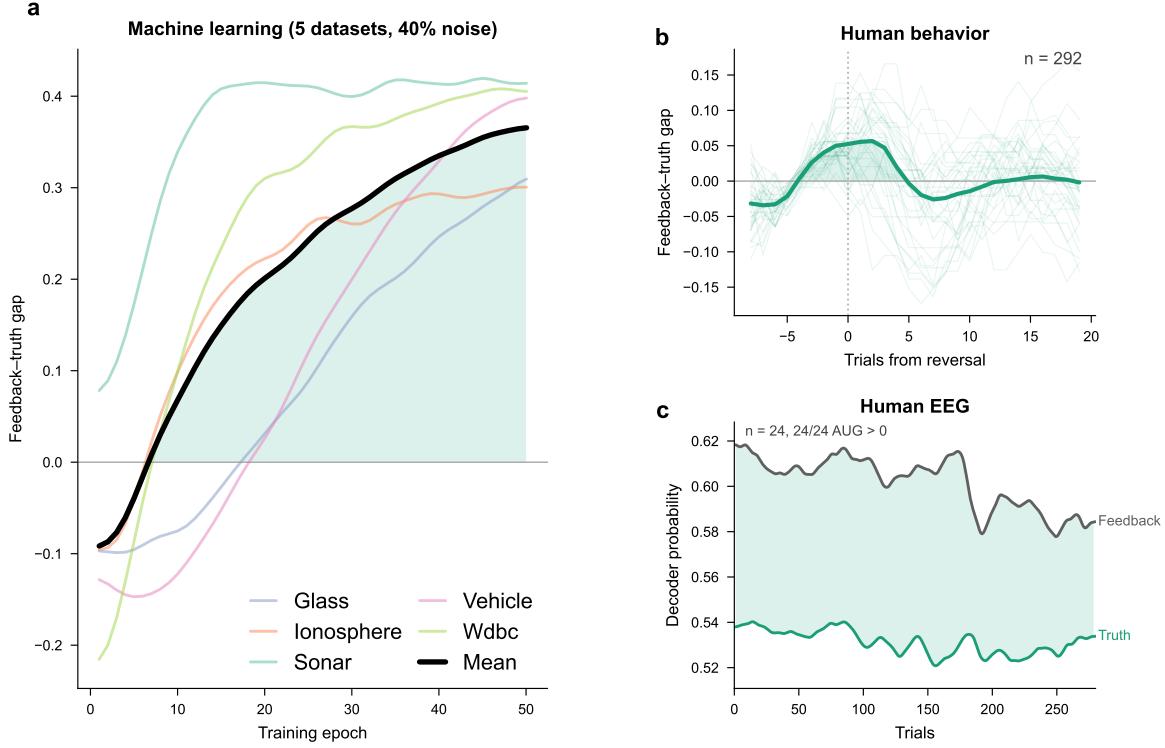
- [1] Arpit, D. *et al.* A Closer Look at Memorization (2017). URL <http://arxiv.org/abs/1706.05394>. ArXiv:1706.05394 [stat].
- [2] Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization (2017). URL <http://arxiv.org/abs/1611.03530>. ArXiv:1611.03530 [cs].
- [3] Frenay, B. & Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **25**, 845–869 (2014). URL <http://ieeexplore.ieee.org/document/6685834/>.
- [4] Song, H., Kim, M., Park, D., Shin, Y. & Lee, J.-G. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **34**, 8135–8153 (2023). URL <https://ieeexplore.ieee.org/document/9729424/>.
- [5] Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006). URL <https://www.nature.com/articles/nature04766>.
- [6] Sutton, R. S. & Barto, A. *Reinforcement learning: an introduction*. Adaptive computation and machine learning (The MIT Press, Cambridge, Massachusetts London, England, 2020), second edition edn.
- [7] Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **95**, 245–258 (2017). URL <https://linkinghub.elsevier.com/retrieve/pii/S0896627317305093>.
- [8] Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**, e253 (2017). URL [https://www.cambridge.org/core/product/identifier/S0140525X16001837/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S0140525X16001837/type/journal_article).
- [9] Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nature Neuroscience* **22**, 1761–1770 (2019). URL <https://www.nature.com/articles/s41593-019-0520-2>.
- [10] Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience* **10** (2016). URL <http://journal.frontiersin.org/Article/10.3389/fncom.2016.00094/abstract>.
- [11] Feldman, V. & Zhang, C. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation (2020). URL <http://arxiv.org/abs/2008.03703>. ArXiv:2008.03703 [cs].
- [12] Toneva, M. *et al.* An Empirical Study of Example Forgetting during Deep Neural Network Learning (2019). URL <http://arxiv.org/abs/1812.05159>. ArXiv:1812.05159 [cs].
- [13] Cools, R., Clark, L., Owen, A. M. & Robbins, T. W. Defining the Neural Mechanisms of Probabilistic Reversal Learning Using Event-Related Functional Magnetic Resonance Imaging. *The Journal of Neuroscience* **22**, 4563–4567 (2002). URL <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.22-11-04563.2002>.

- [14] Izquierdo, A., Brigman, J., Radke, A., Rudebeck, P. & Holmes, A. The neural basis of reversal learning: An updated perspective. *Neuroscience* **345**, 12–26 (2017). URL <https://linkinghub.elsevier.com/retrieve/pii/S030645221600244X>.
- [15] Schoenbaum, G., Roesch, M. R., Stalnaker, T. A. & Takahashi, Y. K. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience* **10**, 885–892 (2009). URL <https://www.nature.com/articles/nrn2753>.
- [16] Clark, L., Cools, R. & Robbins, T. The neuropsychology of ventral prefrontal cortex: Decision-making and reversal learning. *Brain and Cognition* **55**, 41–53 (2004). URL <https://linkinghub.elsevier.com/retrieve/pii/S0278262603002847>.
- [17] Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience* **35**, 1024–1035 (2012). URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1460-9568.2011.07980.x>.
- [18] Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *The Journal of Neuroscience* **32**, 551–562 (2012). URL <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.5498-10.2012>.
- [19] Dayan, P. & Balleine, B. W. Reward, Motivation, and Reinforcement Learning. *Neuron* **36**, 285–298 (2002). URL <https://linkinghub.elsevier.com/retrieve/pii/S0896627302009637>.
- [20] Cavanagh, J. F. & Frank, M. J. Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences* **18**, 414–421 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S1364661314001077>.
- [21] Ridderinkhof, K. R., Ullsperger, M., Crone, E. A. & Nieuwenhuis, S. The Role of the Medial Frontal Cortex in Cognitive Control. *Science* **306**, 443–447 (2004). URL <https://www.science.org/doi/10.1126/science.1100301>.
- [22] Polich, J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology* **118**, 2128–2148 (2007). URL <https://linkinghub.elsevier.com/retrieve/pii/S1388245707001897>.
- [23] Holroyd, C. B. & Coles, M. G. H. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review* **109**, 679–709 (2002). URL <https://doi.apa.org/doi/10.1037/0033-295X.109.4.679>.
- [24] Walsh, M. M. & Anderson, J. R. Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews* **36**, 1870–1884 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S0149763412000875>.
- [25] Patrini, G., Rozza, A., Menon, A. K., Nock, R. & Qu, L. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2233–2241 (IEEE, Honolulu, HI, 2017). URL <http://ieeexplore.ieee.org/document/8099723/>.

- [26] Natarajan, N., Dhillon, I. S., Ravikumar, P. & Tewari, A. Learning with noisy labels. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, 1196–1204 (Curran Associates Inc., Red Hook, NY, USA, 2013). Event-place: Lake Tahoe, Nevada.
- [27] Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997). URL <https://www.science.org/doi/10.1126/science.275.5306.1593>.
- [28] Schultz, W. Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience* **17**, 183–195 (2016). URL <https://www.nature.com/articles/nrn.2015.26>.
- [29] Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science* **306**, 1940–1943 (2004). URL <https://www.science.org/doi/10.1126/science.1102941>.
- [30] Frank, M. J. & O'Reilly, R. C. A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience* **120**, 497–517 (2006). URL <https://doi.apa.org/doi/10.1037/0735-7044.120.3.497>.
- [31] Cools, R. Dopaminergic modulation of cognitive function—implications for l-DOPA treatment in Parkinson's disease. *Neuroscience & Biobehavioral Reviews* **30**, 1–23 (2006). URL <https://linkinghub.elsevier.com/retrieve/pii/S0149763405000540>.
- [32] Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* **116**, 15849–15854 (2019). URL <https://pnas.org/doi/full/10.1073/pnas.1903070116>.
- [33] Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* **116**, 11537–11546 (2019). URL <https://pnas.org/doi/full/10.1073/pnas.1820226116>.
- [34] Neyshabur, B., Bhojanapalli, S., McAllester, D. & Srebro, N. Exploring Generalization in Deep Learning (2017). URL <http://arxiv.org/abs/1706.08947>. ArXiv:1706.08947 [cs].
- [35] Gershman, S. J. & Daw, N. D. Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology* **68**, 101–128 (2017). URL <https://www.annualreviews.org/doi/10.1146/annurev-psych-122414-033625>.
- [36] Cavanagh, J. F., Masters, S. E., Bath, K. & Frank, M. J. Conflict acts as an implicit cost in reinforcement learning. *Nature Communications* **5**, 5394 (2014). URL <https://www.nature.com/articles/ncomms6394>.
- [37] Casey, B. J. Beyond Simple Models of Self-Control to Circuit-Based Accounts of Adolescent Behavior. *Annual Review of Psychology* **66**, 295–319 (2015). URL <https://www.annualreviews.org/doi/10.1146/annurev-psych-010814-015156>.
- [38] Somerville, L. H. & Casey, B. Developmental neurobiology of cognitive control and motivational systems. *Current Opinion in Neurobiology* **20**, 236–241 (2010). URL <https://linkinghub.elsevier.com/retrieve/pii/S0959438810000073>.

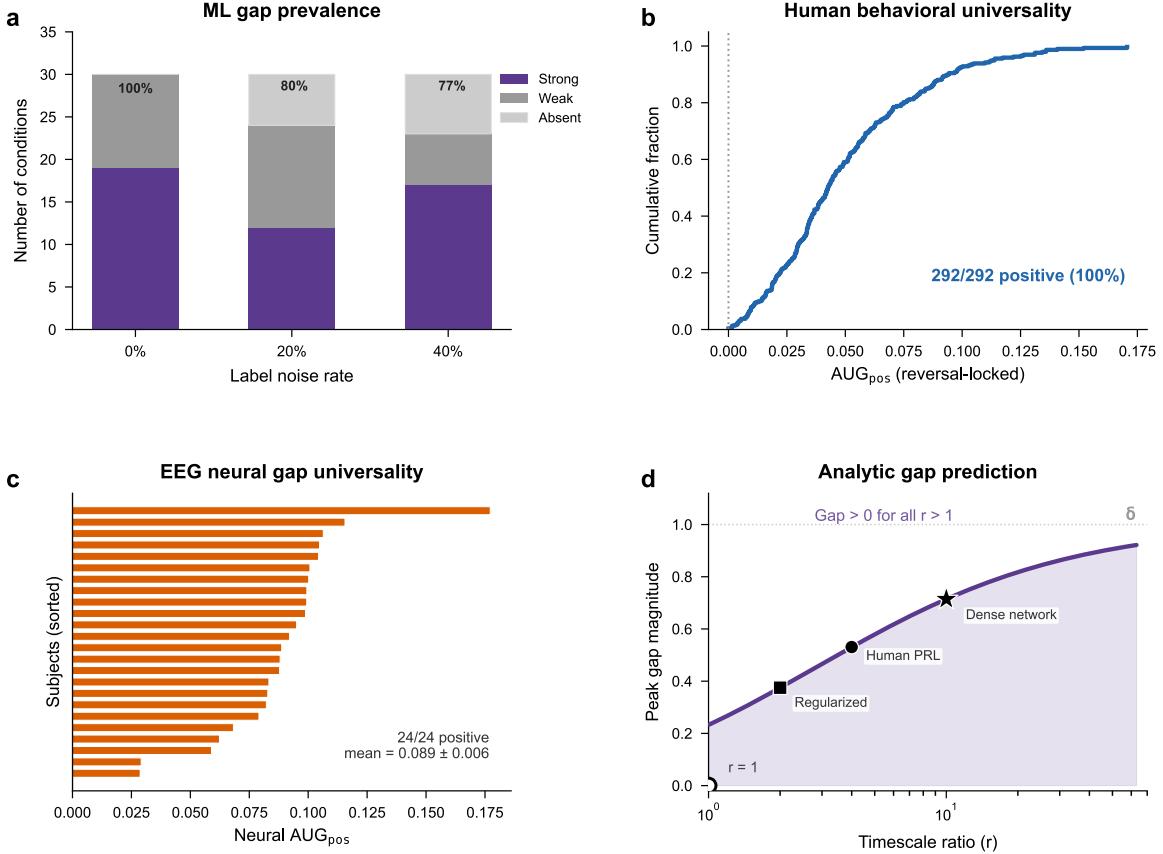
- [39] Hartley, C. A. & Somerville, L. H. The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences* **5**, 108–115 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S2352154615001205>.
- [40] Waltz, J. A. & Gold, J. M. Probabilistic reversal learning impairments in schizophrenia: Further evidence of orbitofrontal dysfunction. *Schizophrenia Research* **93**, 296–303 (2007). URL <https://linkinghub.elsevier.com/retrieve/pii/S092099640700120X>.
- [41] Peterson, D. *et al.* Probabilistic reversal learning is impaired in Parkinson's disease. *Neuroscience* **163**, 1092–1101 (2009). URL <https://linkinghub.elsevier.com/retrieve/pii/S0306452209012068>.
- [42] Maia, T. V. & Frank, M. J. From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience* **14**, 154–162 (2011). URL <https://www.nature.com/articles/nn.2723>.
- [43] Murray, G. K. *et al.* Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Molecular Psychiatry* **13**, 267–276 (2008). URL <https://www.nature.com/articles/4002058>.
- [44] Waltz, J. A. *et al.* Patients with Schizophrenia have a Reduced Neural Response to Both Unpredictable and Predictable Primary Reinforcers. *Neuropsychopharmacology* **34**, 1567–1577 (2009). URL <https://www.nature.com/articles/npp2008214>.
- [45] Rendell, L. *et al.* Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science* **328**, 208–213 (2010). URL <https://www.science.org/doi/10.1126/science.1184719>.
- [46] Richerson, P. J. & Boyd, R. *Culture and the evolutionary process* (University of Chicago press, Chicago London, 1985).
- [47] Rescorla, R. & Wagner, A. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory*, vol. Vol. 2 (1972). Journal Abbreviation: Classical Conditioning II: Current Research and Theory.
- [48] Siegel, S. & Allan, L. G. The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review* **3**, 314–321 (1996). URL <http://link.springer.com/10.3758/BF03210755>.
- [49] Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L. & Collins, A. G. Reinforcement learning and Bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience* **55**, 101106 (2022). URL <https://linkinghub.elsevier.com/retrieve/pii/S1878929322000494>.
- [50] Stoltz, C., Pickering, A. & Mueller, E. M. Reward gain and punishment avoidance reversal learning (2022). URL <https://openneuro.org/datasets/ds004295/versions/1.0.0>.

**Figure 1 | The feedback–truth gap appears across machines and humans.**



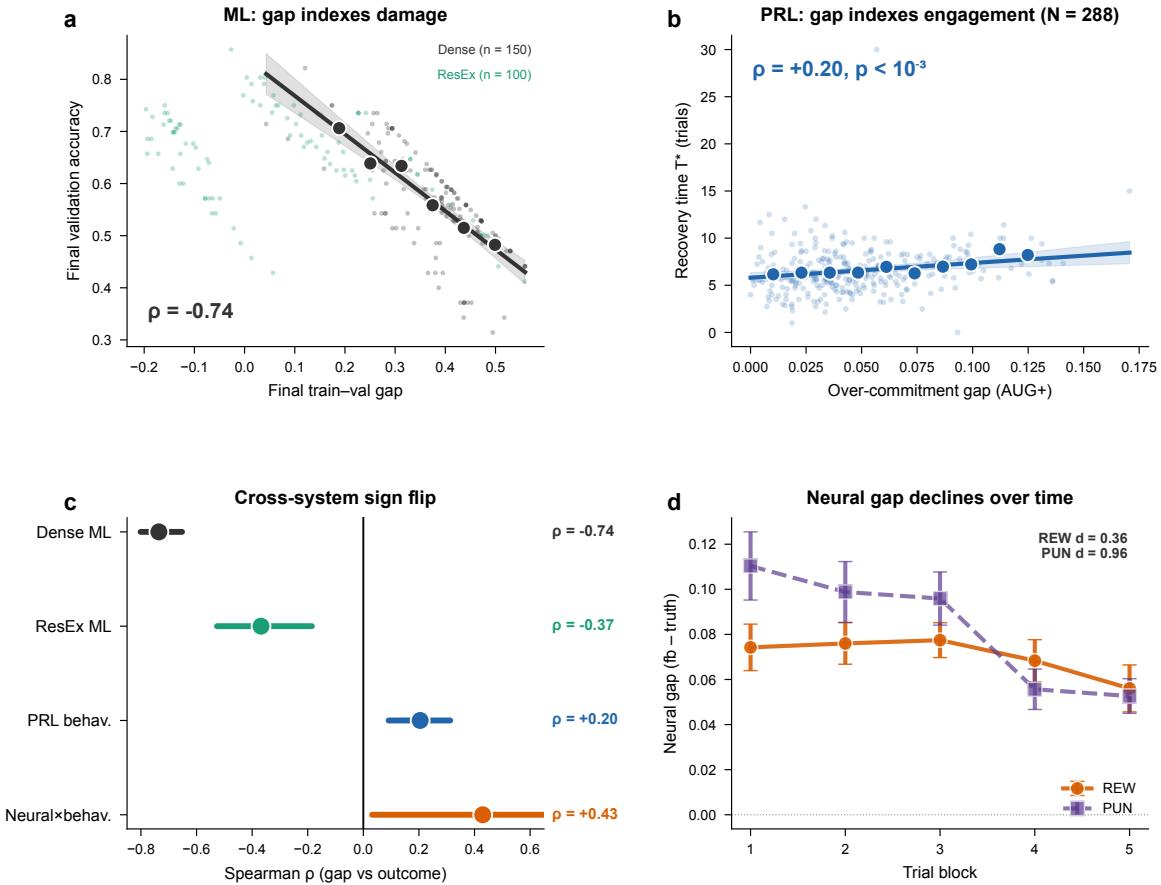
Each panel plots feedback-aligned minus truth-aligned performance in a system that receives noisy supervision. All three systems produce the gap; only the temporal trajectory differs. **a**, An unregularized dense neural network trained on tabular data with 40% label noise (representative run at the 46th percentile of gap magnitude across  $n = 90$  model-noise configurations; Ionosphere dataset; see Extended Data Fig. 1). Training accuracy (grey) exceeds validation accuracy (teal), and the resulting gap (shaded) persists and grows throughout training with no sign of self-correction. **b**, Human reversal learning ( $N = 292$  subjects, 1,764 reversal epochs from the Eckstein *et al.* dataset; ref. 11). The baseline-corrected feedback–truth gap  $\Delta\text{gap}(t)$  is aligned to reversal ( $t = 0$ ). A transient positive deflection (over-commitment) appears immediately post-reversal and returns toward baseline within  $\sim 5$ –10 trials (8-trial rolling window; baseline = mean pre-reversal gap). **c**, Decoder-based neural gap during reward learning ( $n = 24$  subjects from the Stolz *et al.* EEG dataset; ref. 12). Logistic regression decoders trained on theta power, beta power, and P300 amplitude yield  $P(\text{feedback}+|\text{EEG})$  (grey) and  $P(\text{expectation}+|\text{EEG})$  (teal), where expectation = pre-feedback confidence in the chosen option. The gap between decoder probabilities diminishes across trials (15-trial rolling mean), consistent with progressive neural adaptation; the shading transition marks the onset of gap narrowing. Color convention for all panels: teal = truth, grey = feedback, shaded = learning gap.

**Figure 2 |** The feedback–truth gap is universal across systems and mathematically inevitable.



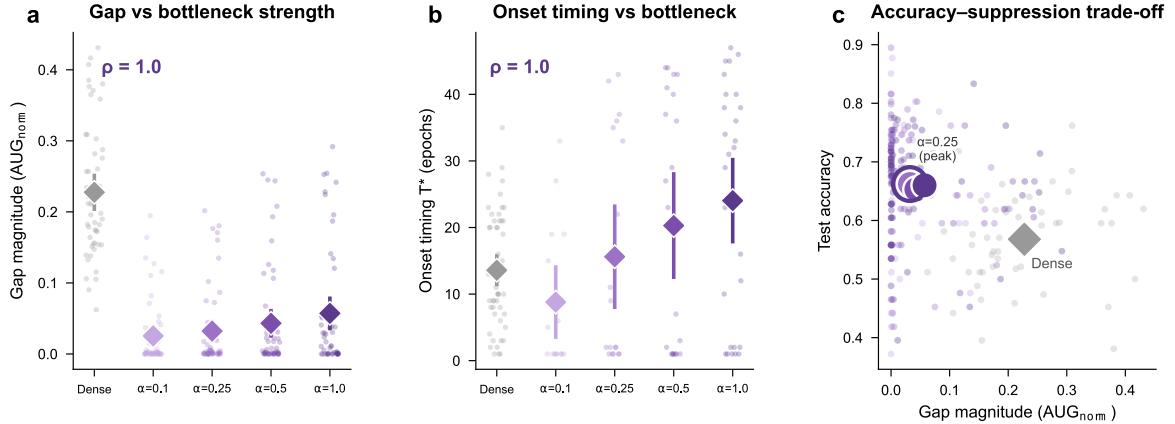
Gap prevalence across machines and humans, together with the two-timescale model that predicts it. **a**, ML prevalence across 90 conditions (5 datasets  $\times$  6 architectures  $\times$  3 noise rates). Stacked bars classify each condition as strong gap (dark purple,  $\text{AUG} \geq 0.05$ ), weak (light purple,  $0.005 \leq \text{AUG} < 0.05$ ), or absent (grey,  $\text{AUG} < 0.005$ ). Across all 90 conditions, 77 (86%) show a measurable gap. Gap prevalence increases with noise rate and exceeds 90% at 40% noise. **b**, Cumulative distribution of over-commitment gap ( $\text{AUG}_{\text{pos}}$ ) across  $N = 292$  PRL subjects. Every subject shows gap  $> 0$  (292/292, 100%). Vertical dashed lines mark quartiles; median  $\text{AUG}_{\text{pos}} = 0.043$ . **c**, Per-subject neural gap ( $\text{AUG}_{\text{pos}}$  fraction, feedback dominance) in the reward task ( $N = 22$  after class-balance exclusion; see Methods and Extended Data Fig. 9), sorted by magnitude. 21/22 subjects (95%) show feedback-dominant neural representation (mean  $\pm$  SEM =  $0.82 \pm 0.03$ ). Orange bars are individual subjects; white dashed line marks 50% (parity). **d**, Peak gap magnitude from the closed-form solution plotted against timescale ratio  $r = \alpha_{\text{fast}}/\alpha_{\text{slow}}$  (log scale). Purple curve:  $\text{gap}_{\text{max}} = \delta \cdot r^{-1/(r-1)} \cdot (1 - r^{-1})$ . The gap is positive for all  $r > 1$  and equals zero exactly at  $r = 1$  (matched timescales, black circle). Empirical systems overlaid: Dense network (star), Human PRL (circle), Regularized network (square). See Extended Data Fig. 2c for the full phase diagram.

**Figure 3 |** The gap indexes damage in unregulated networks but structured engagement in regulated humans.



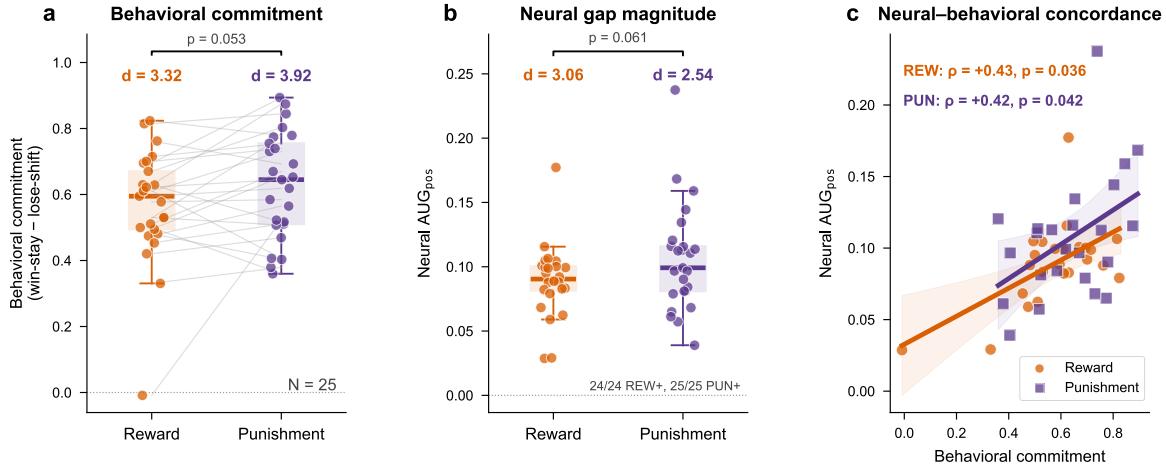
Whether the gap predicts harm or engagement depends on whether the system regulates it. **a**, In unregulated dense networks, a larger final train-validation gap predicts lower validation accuracy (150 dense runs, black; 100 sparse-residual runs, teal). Dense networks: Spearman  $\rho = -0.74$ . Regression line with 95% CI shown for dense networks. **b**, In human PRL the relationship reverses ( $N = 288$  with sufficient post-reversal data). Larger over-commitment (AUG<sub>pos</sub>) predicts faster recovery ( $T^*$  in trials; Spearman  $\rho = +0.20$ ,  $p < 10^{-3}$ ), suggesting that the gap reflects contingency engagement rather than damage. Regression line with 95% CI; point sizes proportional to density. **c**, Forest plot of gap–outcome Spearman correlations across four systems. Dense ML:  $\rho = -0.74$  (harm); ResEx ML:  $\rho = -0.37$  (attenuated); PRL behavioral:  $\rho = +0.20$  (sign reversal); Neural × behavioral:  $\rho = +0.43$  (concordant amplification). 95% CIs shown. **d**, Mean neural gap (feedback minus truth decoder probability) across five trial blocks. Reward (orange):  $d = 0.36$ ; punishment (purple):  $d = 0.82$ . Error bars:  $\pm 1$  SEM. The declining trajectory indicates progressive neural recalibration toward truth, distinct from the persistent accumulation seen in networks and the behavioral recovery observed in reversal learning.

**Figure 4 | The gap is causally tunable via architectural intervention.**



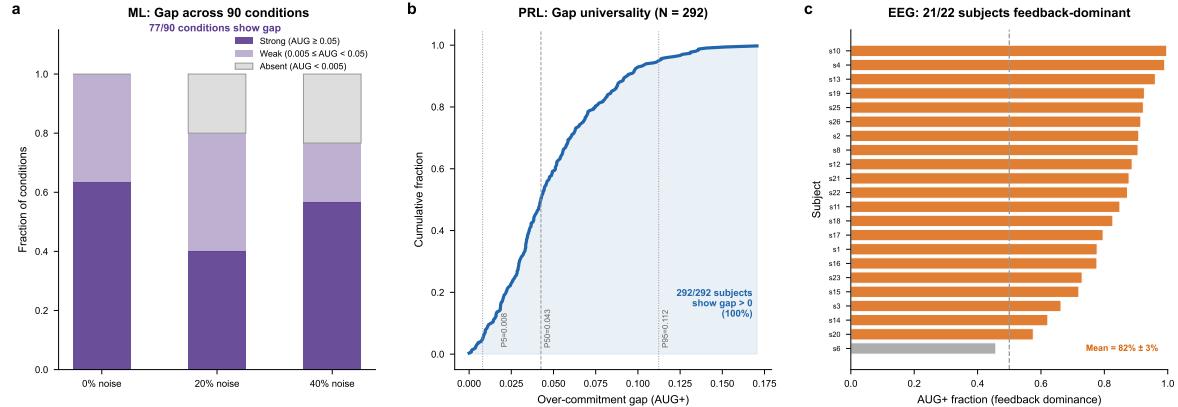
The sparse-residual bottleneck parameter  $\alpha$  provides graded, monotonic control over memorization dynamics and reveals a dissociation between gap suppression and generalization. **a**, Normalized cumulative gap ( $AUG_{norm}$ ) increases monotonically with  $\alpha$  (Spearman  $\rho = 1.0$  across all datasets with measurable gaps). The dense baseline (grey diamond) shows the largest gap. Light points are individual runs; diamonds are group means with 95% CI. Five datasets, 40% label noise, 10 seeds per configuration. **b**, Gap onset time ( $T^*$  in epochs) also increases monotonically with  $\alpha$  ( $\rho = 1.0$ ): more constrained architectures delay memorization. Dense baseline (grey diamond) shows earliest onset. Error bars: 95% CI. **c**, Test accuracy plotted against gap magnitude ( $AUG_{norm}$ ). The relationship is non-monotonic – accuracy peaks at  $\alpha = 0.25$  (teal), not at maximal gap suppression. At  $\alpha = 1.0$  (dark teal), accuracy falls below the moderate-suppression level, exposing a capacity–memorization trade-off. Dense baseline (grey diamond) achieves lower accuracy with larger gap. Point sizes proportional to sample count; error bars: 95% CI.

**Figure 5 | Neural-behavioral concordance in the feedback-truth gap.**



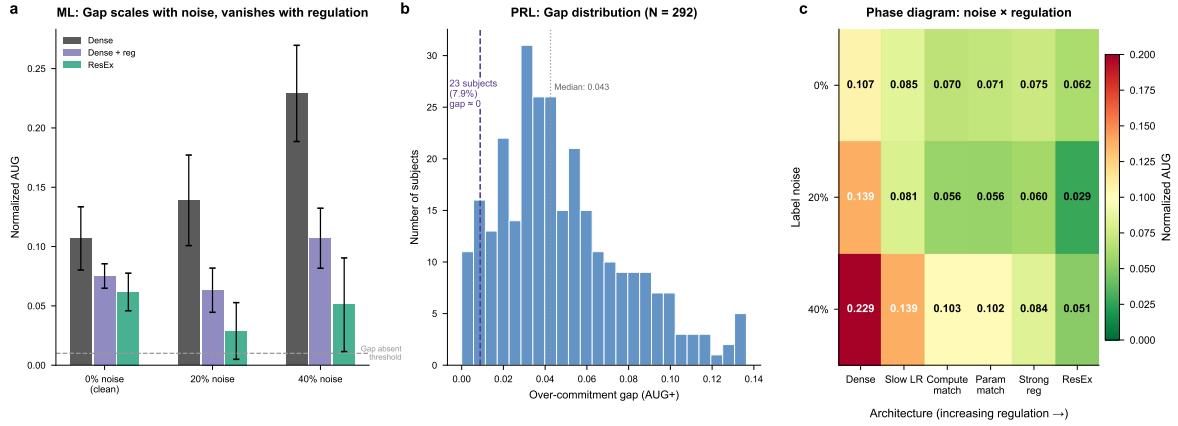
EEG dataset ( $N = 25$ ; ref. 12). The roughly 10-fold amplification from neural representation to overt behavior is the central observation. **a**, Behavioral commitment (win-stay minus lose-shift) for reward (orange,  $d = 3.32$ ) and punishment (purple,  $d = 3.92$ ). Every subject shows positive commitment in both conditions. Connected lines are paired within-subject values; box plots show median and IQR. Difference between tasks:  $d = 0.40$ ,  $p = 0.053$  (two-sided paired  $t$ -test). **b**, Neural gap magnitude (AUG<sub>pos</sub> from EEG decoders) is roughly an order of magnitude smaller than the behavioral effect. Reward:  $d = 1.03$  ( $N = 24$ ); punishment:  $d = 1.82$  ( $N = 25$ ). All subjects positive in both conditions (24/24 reward, 25/25 punishment). Box plots with individual data; error bars:  $\pm 1$  SEM. Difference:  $d = 0.39$ ,  $p = 0.061$ . **c**, Subjects with larger neural over-representation also show stronger behavioral commitment. Reward: Spearman  $\rho = +0.43$ ,  $p = 0.036$  ( $n = 24$ ); punishment:  $\rho = +0.42$ ,  $p = 0.042$  ( $n = 24$ ). Regression lines with 95% CI shown per task. Orange circles: reward; purple squares: punishment.

## Extended Data Figure 1 | Gap emergence rates across systems.



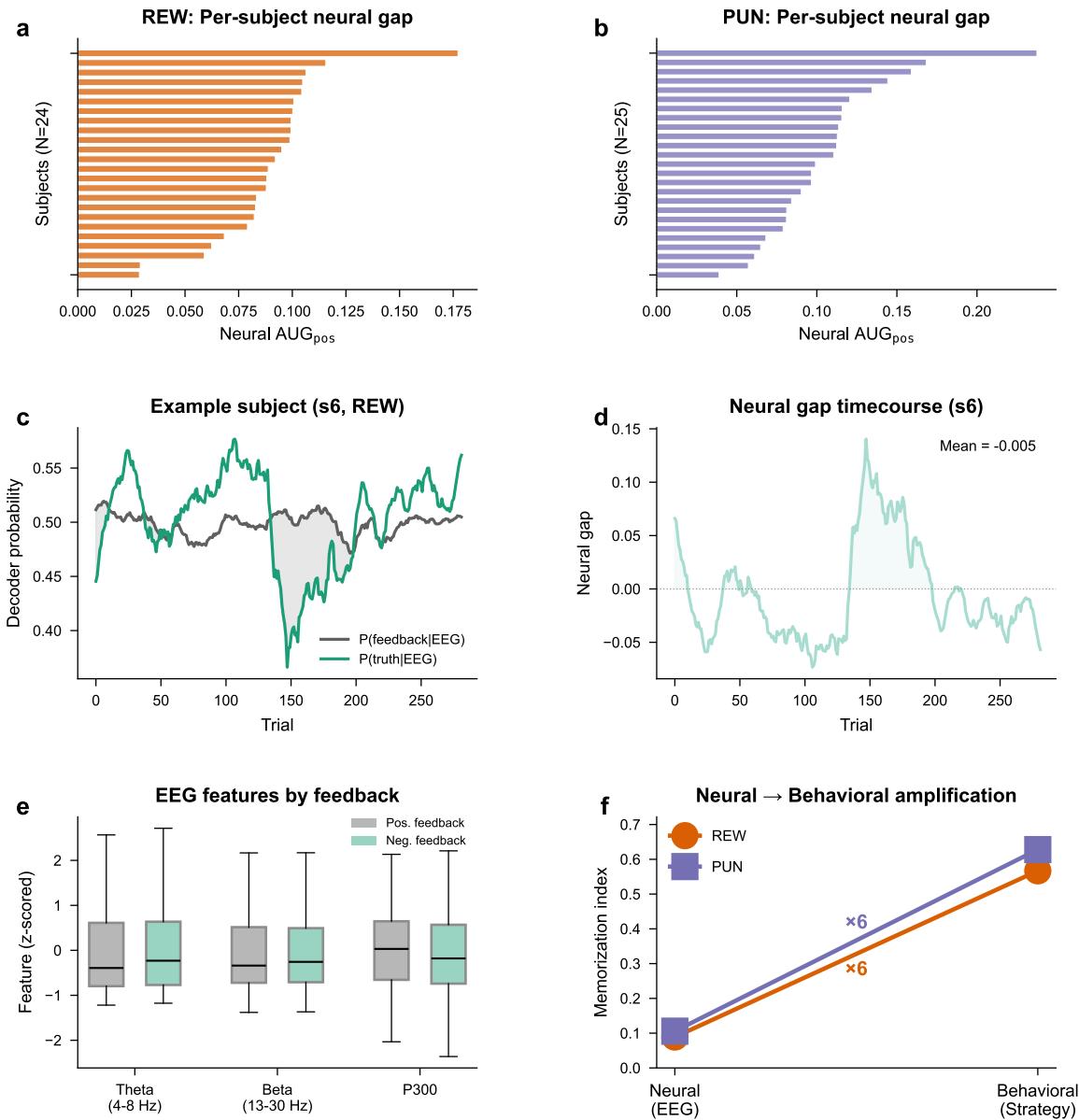
Under noisy conditions the gap is near-universal. Gap = feedback-aligned minus truth-aligned performance throughout. **a**, Fraction of ML conditions showing gap ( $AUG \geq 0.005$ ) at each noise rate (5 datasets  $\times$  6 architectures  $\times$  3 noise rates = 90 conditions). Across all 90 conditions, 77 (86%) show a measurable gap; the proportion exceeds 90% at 40% noise. Bars are proportions; no error bars (each condition is a single count). **b**, Cumulative distribution of  $AUG_{pos}$  across  $N = 292$  PRL subjects (Eckstein dataset; ref. 11). Virtually all subjects show positive over-commitment (median  $AUG_{pos} = 0.043$ ). Feedback = reward received; truth = objectively better option. **c**, Per-subject  $AUG+$  fraction (feedback dominance;  $N = 22$  non-excluded subjects from Stoltz EEG dataset; ref. 12). 95% of subjects exceed the 50% parity line (mean  $\pm$  SEM =  $0.82 \pm 0.03$ ). Feedback = reinforcer valence; truth = pre-feedback expectation.

## Extended Data Figure 2 | Boundary conditions: when the gap vanishes.



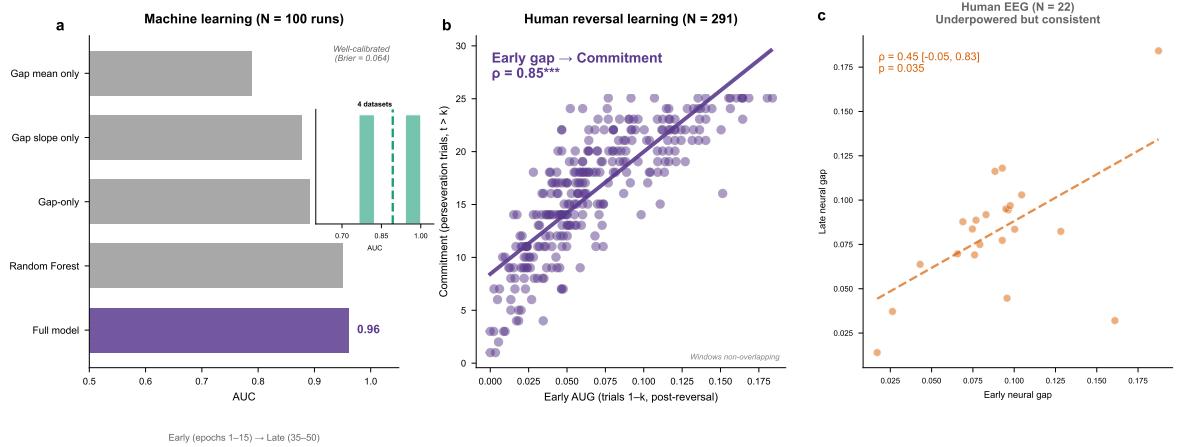
The inevitability claim is falsifiable: the gap disappears under specific counterfactual conditions. Gap = feedback-aligned minus truth-aligned performance throughout. **a**, AUG by noise rate and architecture across 90 ML conditions (5 datasets  $\times$  6 architectures  $\times$  3 noise rates). The gap scales with noise and vanishes under strong regularization or 0% noise; 15/90 conditions show near-zero gap (AUG < 0.01). Points are individual conditions. **b**, Distribution of AUG<sub>pos</sub> across  $N = 292$  PRL subjects. 23 subjects (7.9%) show near-zero gap, consistent with low-noise or matched-timescale subgroups. Feedback = reward received; truth = objectively better option. **c**, Phase diagram showing AUG across noise rate  $\times$  architecture. Color scale is mean normalized AUG; cells with white squares have AUG < 0.01 (gap absent). The gap vanishes under clean labels, strong regularization, or capacity suppression.

### Extended Data Figure 3 | Neural validation details.



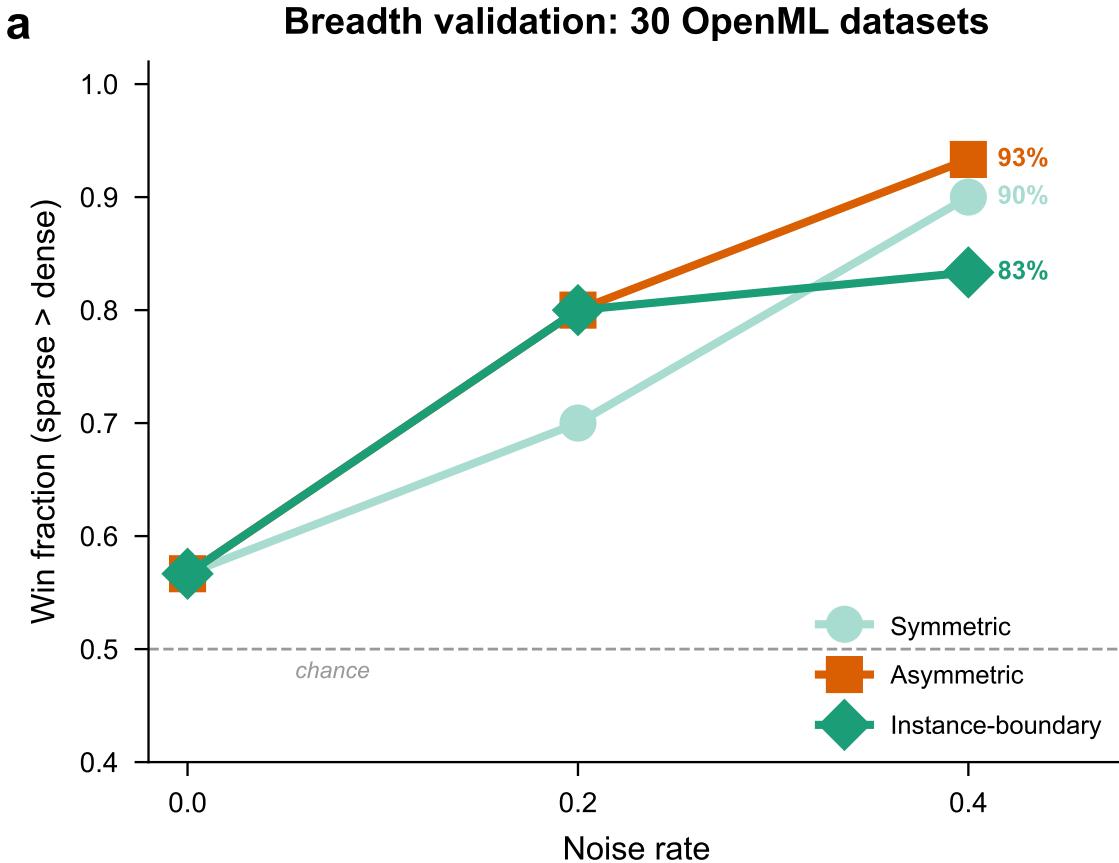
Subject-level neural gap data, decoder validation, and cross-modal dynamics. Feedback = reinforcer valence (reward or punishment); truth = prior expectation (pre-feedback confidence; see Methods). **a–b**, Individual neural AUG<sub>pos</sub> values for reward ( $N = 24$ ) and punishment ( $N = 25$ ), sorted by magnitude. All values are positive. Bars are individual subjects. **c–d**, An example subject (s6, reward task) illustrates decoder probability trajectories for feedback and expectation (15-trial rolling mean). Shading marks the gap between decoder curves. **e**, Distributions of the three EEG features by feedback type: theta power, beta power, and P300 amplitude. Box plots show median and IQR; whiskers extend to  $1.5 \times$  IQR. **f**, The memorization index is approximately 10-fold larger at the behavioral than neural level (in standardized effect size), illustrating the amplification from cortical representation to overt strategy. Error bars:  $\pm 1$  SEM.

## Extended Data Figure 4 | Early training dynamics predict later memorization regime.



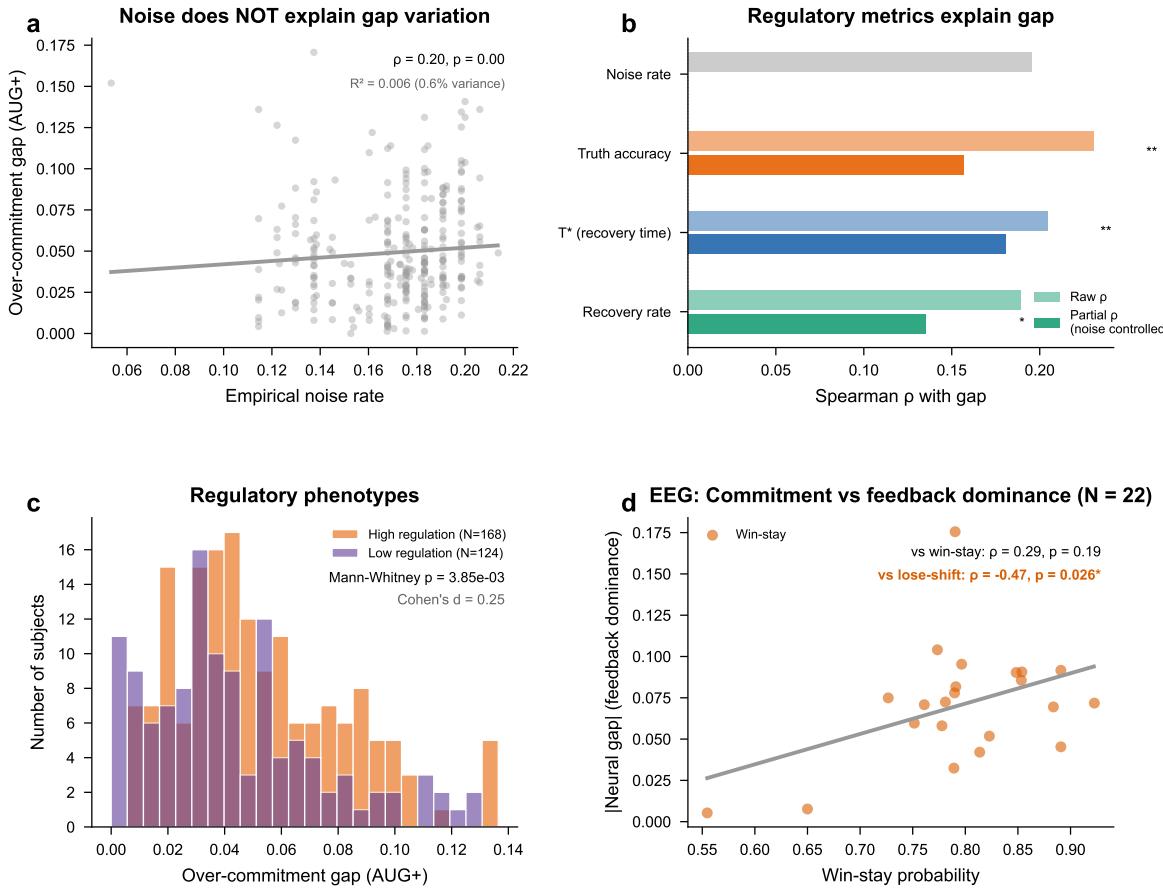
Features measured early in training (epochs 1–15) predict the later memorization regime (epochs 35+) with a strict 20-epoch temporal buffer, indicating that gap dynamics are deterministic rather than stochastic. **a**, Cross-validated AUC across  $N = 100$  ML runs (5 datasets, 40% noise). Logistic regression on 14 early features achieves  $AUC = 0.96$  (Brier score = 0.064); the single best feature (gap\_slope) alone reaches  $AUC = 0.88$ . Bars are AUC; error metric is cross-validation standard error. Inset: generalization across datasets. **b**, In PRL ( $N = 291$ ; one subject excluded for insufficient post-reversal data), early post-reversal gap predicts later perseveration ( $\rho = 0.85$ ,  $p < 0.001$ ) but not recovery time  $T^*$  ( $\rho = -0.04$ , n.s.). Early pressure and downstream regulation are therefore separable. **c**, In the EEG sample ( $N = 22$ ), early neural gap predicts later signal persistence ( $\rho = 0.45$ ,  $p = 0.035$ ). The sample is small, but the effect is directionally consistent with the behavioral and ML results.

Extended Data Figure 5 | Breadth validation across 30 OpenML datasets.



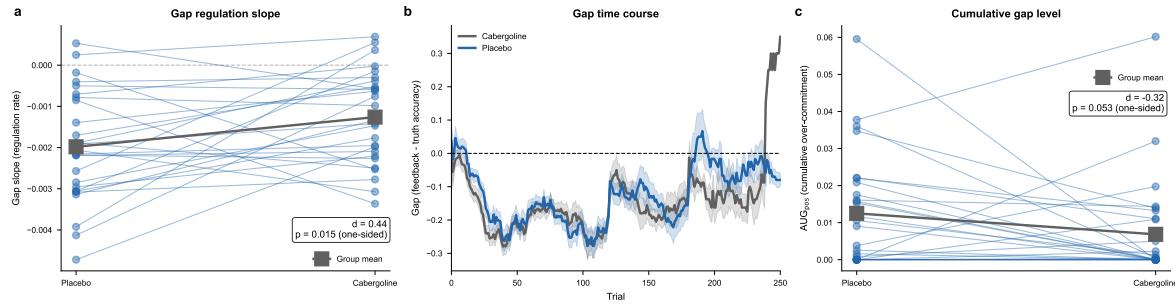
How often the sparse bottleneck outperforms the dense baseline, plotted against noise rate across 30 OpenML datasets ( $n = 2,700$  total runs; 30 datasets  $\times$  3 noise rates  $\times$  10 seeds  $\times$  3 configurations). Truth = held-out test labels (uncorrupted); feedback = noisy training labels. At 0% noise: 57% (no reliable advantage). At 40% noise the win fraction rises to 93% (median  $\Delta = +3.28$  pp). Error bars: 95% binomial CI. The advantage is noise-dependent, not architecture-dominant.

## Extended Data Figure 6 | Individual differences in gap dynamics.



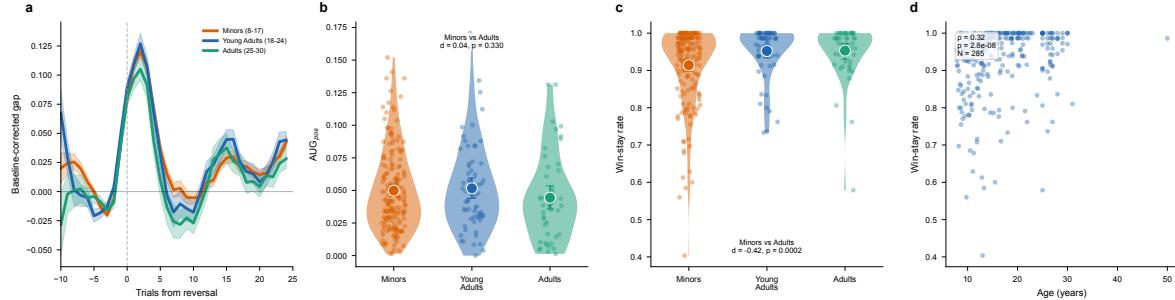
Individual variation in gap magnitude and recovery speed reveals heterogeneity in regulation. Gap = feedback-aligned minus truth-aligned performance; feedback = reward received; truth = objectively better option (PRL) or pre-feedback expectation (EEG). **a**, In PRL ( $N = 292$ ), noise rate positively predicts gap magnitude ( $\rho = 0.20, p = 0.001$ ) but explains only 0.6% of variance – most of the variation lies between individuals, not between task conditions. Each point is one subject; line is a linear fit. **b**, Gap magnitude predicts recovery speed (partial  $\rho = 0.14, p = 0.021$ , controlling for noise rate) and recovery time  $T^*$  (partial  $\rho = 0.18, p = 0.002$ ). PRL,  $N = 292$ . Each point is one subject. **c**, Subjects classified as high-regulation versus low-regulation phenotypes show separable gap profiles ( $d = 0.25$ ). Shading:  $\pm 1$  SEM. **d**, In the EEG sample ( $N = 22$ ), neural AUG<sub>pos</sub> correlates negatively with lose-shift rate ( $\rho = -0.48, p = 0.026$ ): subjects with stronger neural over-representation are less likely to shift after losses. Each point is one subject; line is a linear fit.

## Extended Data Figure 7 | Dopaminergic modulation of the feedback–truth gap.



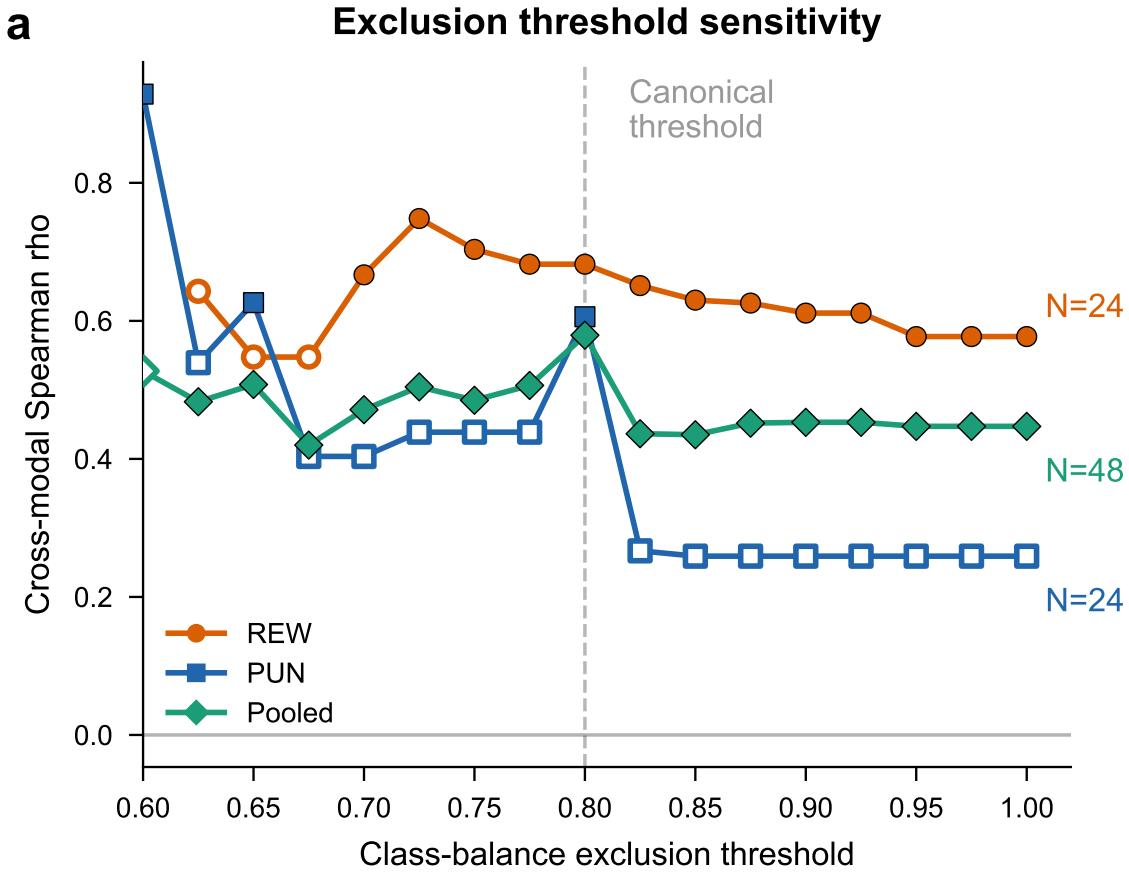
Cabergoline, a dopamine D<sub>2</sub>/D<sub>3</sub> agonist, flattens within-session gap regulation in a double-blind crossover ( $N = 27$ ; Cavanagh & Frank dataset;<sup>36</sup>). Gap = feedback-aligned minus truth-aligned accuracy (20-trial rolling window). **a**, The gap regulation slope (linear trend of gap over trials) is significantly flatter under cabergoline than placebo (one-sided paired *t*-test,  $d = 0.44$ ,  $p = 0.015$ ; Wilcoxon  $p = 0.016$ ). Connected dots are individual subjects; red squares are group means. **b**, Group-averaged gap time courses under cabergoline (red) and placebo (blue), with SEM shading. Under placebo the gap declines more steeply (stronger regulation); cabergoline blunts this trajectory. **c**, Cumulative over-commitment ( $AUG_{pos}$ ) is directionally lower under cabergoline ( $d = -0.32$ , one-sided  $p = 0.053$ ), suggesting that the drug effect extends to both the dynamics and the overall level of the gap.

## Extended Data Figure 8 | Developmental gradient in feedback persistence and gap regulation.



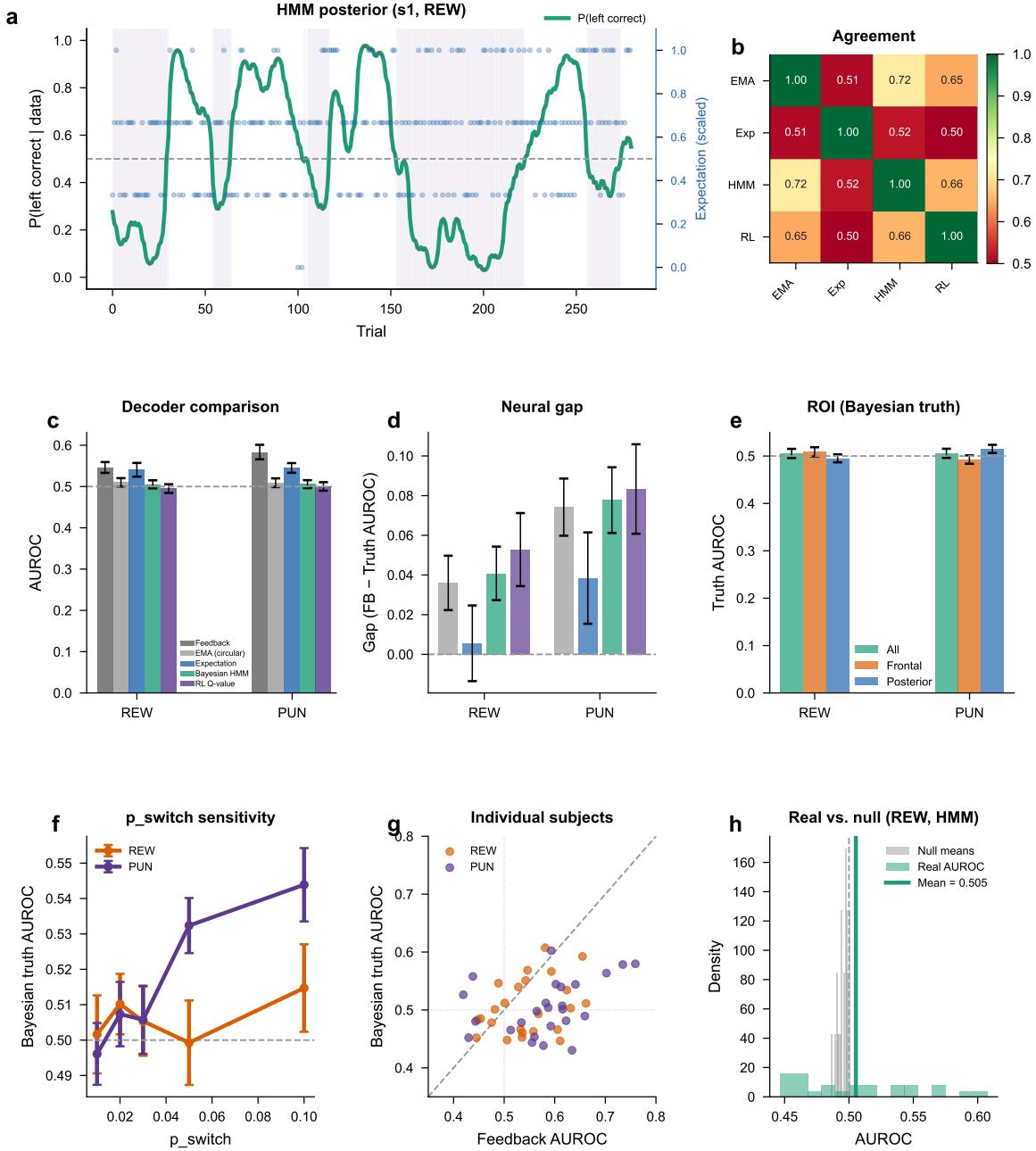
With development, the balance between flexibility and persistence shifts in a direction that parallels the pharmacological effect of dopamine agonism. Gap = feedback-aligned minus truth-aligned accuracy; feedback = reward received; truth = objectively better option. **a**, Reversal-locked gap dynamics for two developmental cohorts ( $sID < 300$ : minors ages 8–17,  $N = 178$ ;  $sID \geq 300$ : adults ages 18–30,  $N = 114$ ; Eckstein *et al.* dataset; ref. 11). Mean baseline-corrected  $\Delta\text{gap}(t)$ , aligned to reversal ( $t = 0$ ). Both groups show a positive post-reversal gap; adults show a slightly flatter recovery trajectory. The faster gap closure in minors reflects reduced feedback persistence rather than enhanced compensatory control. Shading is SEM. **b**, Total over-commitment ( $AUG_{pos}$ ) does not differ significantly between minors and adults ( $d = 0.04, p = 0.66$ , two-sided Mann–Whitney  $U$ ), so the different dynamics in **a** do not translate into a difference in cumulative magnitude. Violin plots with individual data points; error bars are 95% bootstrap CI. **c**, Win-stay rate (probability of repeating a rewarded choice) increases from minors to adults ( $d = -0.42, p < 0.001$ ). The direction parallels the cabergoline effect (Extended Data Fig. 7): both developmental maturation and acute dopamine agonism increase feedback persistence. **d**, Treating age as a continuous variable confirms the pattern (Spearman  $\rho = 0.32, p < 0.0001$ ; age range 7–50 years,  $N = 285$  with matched demographics). The grouped result in **c** is not an artifact of binning.

Extended Data Figure 9 | Cross-modal correlation robustness to exclusion threshold.



The cross-modal Spearman correlation (behavioral commitment versus neural  $AUG_{pos}$ ) is stable across class-balance exclusion thresholds. Subjects whose expectation label imbalance exceeds the threshold are excluded at each step; thresholds from 0.60 to 1.00 were tested. The pooled correlation (reward + punishment) ranges from  $\rho = 0.51$  to  $\rho = 0.59$  and remains significant ( $p < 0.01$ ) throughout. The 0.80 threshold used in the main analysis falls in the middle of this stable range. Per-condition correlations (reward, punishment) are shown separately.  $N$  per threshold varies from 14 to 38 subjects.

**Extended Data Figure 10 | Non-circular truth reconstruction and decoder comparison.**



Post-feedback EEG features decode feedback valence and prior expectation (pre-feedback confidence) but not objective environmental correctness inferred from Bayesian or reinforcement-learning models. We tested four truth definitions: slow exponential moving average of feedback (circular baseline), pre-feedback expectation ratings, Bayesian HMM-inferred correct side (forward-backward with 70/30 emission structure), and fitted RL model  $Q$ -values (M5 parameters from ref. 12). **a**, HMM posterior trajectory for an example subject (s1, REW task):  $P(\text{left correct} \mid \text{all data})$ , with overlaid expectation ratings (scaled). Purple shading marks epochs where the HMM infers right-correct state. **b**, Agreement between truth definitions. Expectation truth is nearly independent of the other three ( $\sim 50\%$  agreement), whereas EMA, HMM, and RL cluster together ( $\sim 65\text{--}72\%$ ). This

is consistent with expectation ratings capturing a genuinely distinct signal. **c**, Decoder AUROC across truth definitions and tasks. Only the expectation-truth decoder significantly exceeds chance (REW: AUROC = 0.541,  $t(23) = 2.43$ ,  $p = 0.012$ ; PUN: AUROC = 0.545,  $t(24) = 3.84$ ,  $p < 10^{-3}$ ). The feedback decoder is shown for reference. **d**, Neural gap (feedback minus truth AUROC) by truth definition. With non-circular expectation truth the gap is near-zero in REW (gap = +0.006,  $p = 0.39$ ) and small in PUN (gap = +0.038,  $p = 0.054$ ), indicating that feedback and expectation representations are nearly balanced in post-feedback EEG. **e**, ROI decomposition for Bayesian truth (all features, frontal-only, posterior-only). No feature subset rescues objective-truth decoding. **f**, Bayesian truth decoder AUROC across  $p_{\text{switch}}$  values (0.01–0.10). The null result for objective truth is stable. **g**, Individual-subject scatter of feedback AUROC versus Bayesian truth AUROC. The null is not driven by outlier subjects. **h**, Real versus null AUROC distributions for the Bayesian truth decoder (REW task). HMM validation: reward rate when HMM-inferred correct = 0.86, when incorrect = 0.28, confirming reconstruction validity against the known 70/30 contingency structure.

**Extended Data Table 1 | Gap dynamics across noise-robust training methods.**

Method	Median AUG <sub>norm</sub>	Median $T^*$	Frac. AUG > 0.005	Median val. acc.
Dense (baseline)	0.121	24.0	73%	0.646
Sparse-residual	0.000	50.0 (censored)	30%	0.770
Co-teaching	0.000	50.0 (censored)	23%	0.806
Forward loss correction	0.009	48.8	57%	0.705

Four training methods applied to 30 OpenML datasets with 40% symmetric label noise (3 seeds each). Median AUG<sub>norm</sub> and  $T^*$  computed per dataset (mean across seeds), then median across datasets. All methods produce a measurable gap in a substantial fraction of datasets; no method eliminates it entirely. The gap is suppressed but not abolished by algorithmic intervention, consistent with inevitability under the two-timescale framework.

**Extended Data Table 2 | Sparse-residual gap suppression across 30 datasets.**

Noise rate	$N$ datasets	Median $\Delta\text{AUG}$	Frac. $\Delta\text{AUG} < 0$	Median $\Delta_{\text{acc}}$	Frac. $\Delta_{\text{acc}} > 0$	$T^*$ delay frac.
0.2	30	-0.055	80% [63–93%]	+0.052	87% [73–97%]	80%
0.4	30	-0.119	97% [90–100%]	+0.081	93% [83–100%]	91%

Dense baseline versus sparse-residual architecture (hidden dimensions = 256, degree = 5,  $\alpha = 0.25$ ) across 30 OpenML datasets at two noise rates (3 noise types  $\times$  10 seeds = 18 runs per dataset per architecture).  $\Delta\text{AUG}$  = sparse-residual minus dense (negative = suppression).  $\Delta_{\text{acc}}$  = sparse-residual minus dense (positive = improvement).  $T^*$  delay fraction: proportion of datasets where sparse-residual delays the onset of memorization. Sparse-residual models use approximately 5% of dense-baseline compute (median MACs ratio = 0.052).

**Extended Data Table 3 | Topology null study.**

Metric	Value
$N$ datasets	10
Median $\Delta\text{AUG}_{\text{norm}}$ (random – expander)	0.000
Frac. $ \Delta\text{AUG}_{\text{norm}}  < 0.01$	100%
Median $\Delta_{\text{test acc}}$	-0.0004
Frac. ordering consistent (random $\geq$ expander)	0%

Expander-graph versus random-regular graph wiring within the same sparse-residual scaffold (10 datasets, symmetric noise at 40%, 3 seeds).  $\Delta\text{AUG}_{\text{norm}}$  = random-regular minus expander. Graph topology produces no detectable difference in gap dynamics: median  $\Delta\text{AUG}_{\text{norm}} = 0.000$  and 100% of datasets fall within  $|\Delta| < 0.01$ . The sparse-residual scaffolding, not specific graph structure, is the operative ingredient for gap suppression.

**Extended Data Table 4 | Reinforcement-learning parameters selectively predict gap dynamics.**

Outcome	Predictor	$\rho$	$p$	Residualized $\rho$	Residualized $p$
AUG <sub>pos</sub>	$\alpha$ (standard RW)	0.291	$< 10^{-6}$	0.057	0.328
AUG <sub>pos</sub>	$\pi$ (perseveration)	-0.264	$< 10^{-5}$	-0.083	0.159
AUG <sub>pos</sub>	$\alpha^-$ (asymmetric)	0.291	$< 10^{-6}$	-	-
AUG <sub>pos</sub>	$\alpha^+ - \alpha^-$ (asymmetry)	-0.222	$< 10^{-4}$	-	-
Gap slope	$\alpha$ (standard RW)	0.333	$< 10^{-8}$	0.367	$< 10^{-10}$
Gap slope	$\pi$ (perseveration)	-0.149	0.011	-0.207	$< 10^{-3}$
Gap slope	$\alpha^-$ (asymmetric)	0.314	$< 10^{-7}$	-	-
Gap slope	$\alpha^+ - \alpha^-$ (asymmetry)	-0.129	0.027	-0.113	0.054

Spearman correlations between fitted RL parameters and gap metrics ( $N = 292$  subjects, Eckstein 2022 dataset). Three models fitted: standard Rescorla–Wagner (RW), RW with perseveration ( $\pi$ ), and asymmetric RW ( $\alpha^+$ ,  $\alpha^-$ ). The standard learning rate  $\alpha$  predicts both gap magnitude (AUG<sub>pos</sub>) and recovery speed (gap slope), consistent with the two-timescale framework. Perseveration  $\pi$  shows inverse relationships. These associations survive residualization for task-level confounds (noise rate, trial count, reversal count).

# Supplementary Information

Learning under noisy supervision is governed by a feedback–truth gap

Elan Schonfeld<sup>1</sup>      Elias Wisnia<sup>2</sup>

## Contents

<b>1 Supplementary Derivation: Two-Timescale Gap</b>	<b>2</b>
1.1 Model definition . . . . .	2
1.2 Gap dynamics after a state change . . . . .	2
1.3 Proof of strict positivity . . . . .	2
1.4 Peak timing . . . . .	2
1.5 Peak magnitude . . . . .	3
1.6 Limiting behavior . . . . .	3
1.7 Cumulative gap . . . . .	3
<b>2 Supplementary Methods</b>	<b>4</b>
2.1 Machine learning: extended experimental details . . . . .	4
2.1.1 Dataset specifications . . . . .	4
2.1.2 Architecture details . . . . .	4
2.1.3 Memorization metric definitions . . . . .	4
2.1.4 Alpha-grid causal intervention . . . . .	5
2.2 Human probabilistic reversal learning: extended methods . . . . .	5
2.2.1 Dataset and participants . . . . .	5
2.2.2 Task structure . . . . .	5
2.2.3 Reversal-locked analysis . . . . .	5
2.2.4 AUG <sub>pos</sub> computation . . . . .	5
2.3 Human reward/punishment learning with EEG: extended methods . . . . .	6
2.3.1 Dataset and participants . . . . .	6
2.3.2 EEG preprocessing . . . . .	6
2.3.3 Neural features . . . . .	6
2.3.4 Neural gap decoder . . . . .	6
2.3.5 Source localization . . . . .	7
2.3.6 Cross-modal correlation . . . . .	7
<b>3 Supplementary Results</b>	<b>8</b>
3.1 Complete statistical summary . . . . .	8
3.2 Reinforcement learning model parameters . . . . .	8
3.3 Positioning relative to noise-robust machine learning . . . . .	8
<b>4 Supplementary Figures</b>	<b>10</b>

# 1 Supplementary Derivation: Two-Timescale Gap

Below we work out the gap in closed form for the two-timescale learner, show that it is strictly positive whenever the timescales differ, and derive the peak magnitude.

## 1.1 Model definition

A learner tracks an environment state  $s \in \{0, 1\}$  via two channels, each updated by an exponential moving average:

$$x_{\text{fast}}(t+1) = x_{\text{fast}}(t) + \alpha_{\text{fast}} \cdot (o(t) - x_{\text{fast}}(t)) \quad (1)$$

$$x_{\text{slow}}(t+1) = x_{\text{slow}}(t) + \alpha_{\text{slow}} \cdot (s(t) - x_{\text{slow}}(t)) \quad (2)$$

where  $o(t)$  is the observed feedback signal (equal to  $s(t)$  with probability  $1 - \varepsilon$  and  $1 - s(t)$  with probability  $\varepsilon$ ) and  $0 < \alpha_{\text{slow}} < \alpha_{\text{fast}} \leq 1$ . In the continuous-time limit, these become exponential decay processes with rates  $\alpha_{\text{fast}}$  and  $\alpha_{\text{slow}}$  respectively.

## 1.2 Gap dynamics after a state change

Consider a state reversal of magnitude  $\Delta$  at time  $t = 0$ . Both channels begin with the same initial displacement  $\Delta$  from the new true state. In expectation, the residual error of each channel decays exponentially:

$$\mathbb{E}[\text{error}_{\text{fast}}(t)] = \Delta \cdot e^{-\alpha_{\text{fast}} \cdot t} \quad (3)$$

$$\mathbb{E}[\text{error}_{\text{slow}}(t)] = \Delta \cdot e^{-\alpha_{\text{slow}} \cdot t} \quad (4)$$

The feedback-truth gap is the difference in alignment between the two channels (equivalently, the difference in residual errors):

$$\text{gap}(t) = \Delta \cdot (e^{-\alpha_{\text{slow}} \cdot t} - e^{-\alpha_{\text{fast}} \cdot t}) \quad (5)$$

## 1.3 Proof of strict positivity

**Proposition 1.** *If  $\alpha_{\text{fast}} > \alpha_{\text{slow}} > 0$ , then  $\text{gap}(t) > 0$  for all  $t > 0$ .*

*Proof.* Since  $\alpha_{\text{fast}} > \alpha_{\text{slow}}$ , we have  $\alpha_{\text{fast}} \cdot t > \alpha_{\text{slow}} \cdot t$  for all  $t > 0$ . Because  $f(x) = e^{-x}$  is strictly decreasing,  $e^{-\alpha_{\text{fast}} \cdot t} < e^{-\alpha_{\text{slow}} \cdot t}$ . Therefore  $e^{-\alpha_{\text{slow}} \cdot t} - e^{-\alpha_{\text{fast}} \cdot t} > 0$ , and since  $\Delta > 0$ ,  $\text{gap}(t) > 0$ .  $\square$

**Corollary (Counterfactual).** When  $\alpha_{\text{fast}} = \alpha_{\text{slow}}$ ,  $\text{gap}(t) = 0$  for all  $t$ , regardless of noise level  $\varepsilon$ .

## 1.4 Peak timing

Differentiating Eq. (5) with respect to  $t$  and setting the derivative to zero:

$$\frac{d}{dt} \text{gap}(t) = \Delta \cdot (-\alpha_{\text{slow}} \cdot e^{-\alpha_{\text{slow}} \cdot t} + \alpha_{\text{fast}} \cdot e^{-\alpha_{\text{fast}} \cdot t}) = 0 \quad (6)$$

This yields:

$$\alpha_{\text{fast}} \cdot e^{-\alpha_{\text{fast}} \cdot t^*} = \alpha_{\text{slow}} \cdot e^{-\alpha_{\text{slow}} \cdot t^*} \quad (7)$$

Taking logarithms:

$$\ln\left(\frac{\alpha_{\text{fast}}}{\alpha_{\text{slow}}}\right) = (\alpha_{\text{fast}} - \alpha_{\text{slow}}) \cdot t^* \quad (8)$$

Defining the timescale ratio  $r = \alpha_{\text{fast}}/\alpha_{\text{slow}}$ :

$$t^* = \frac{\ln r}{\alpha_{\text{slow}} \cdot (r - 1)} \quad (9)$$

## 1.5 Peak magnitude

Substituting  $t^*$  back into Eq. (5):

$$e^{-\alpha_{\text{slow}} \cdot t^*} = e^{-\ln r/(r-1)} = r^{-1/(r-1)} \quad (10)$$

$$e^{-\alpha_{\text{fast}} \cdot t^*} = e^{-r \cdot \ln r/(r-1)} = r^{-r/(r-1)} \quad (11)$$

Therefore:

$$\text{gap}_{\max} = \Delta \cdot (r^{-1/(r-1)} - r^{-r/(r-1)}) \quad (12)$$

This can be simplified by noting that  $r^{-r/(r-1)} = r^{-1/(r-1)} \cdot r^{-1}$ :

$$\text{gap}_{\max} = \Delta \cdot r^{-1/(r-1)} \cdot \left(1 - \frac{1}{r}\right) \quad (13)$$

## 1.6 Limiting behavior

**Large timescale mismatch** ( $r \rightarrow \infty$ ): As  $r \rightarrow \infty$ ,  $r^{-1/(r-1)} \rightarrow 1$  and  $(1 - 1/r) \rightarrow 1$ , so  $\text{gap}_{\max} \rightarrow \Delta$ . The peak gap approaches the full magnitude of the state change.

**Matched timescales** ( $r \rightarrow 1^+$ ): Writing  $r = 1 + \epsilon$  and expanding:

$$r^{-1/(r-1)} = (1 + \epsilon)^{-1/\epsilon} \rightarrow e^{-1} \quad \text{as } \epsilon \rightarrow 0 \quad (14)$$

while  $(1 - 1/r) = \epsilon/(1 + \epsilon) \rightarrow 0$ . Thus  $\text{gap}_{\max} \rightarrow 0$ , confirming that the gap vanishes continuously as timescales converge.

## 1.7 Cumulative gap

The cumulative positive gap, which corresponds to the  $\text{AUG}_{\text{pos}}$  metric used in the main text, integrates Eq. (5) over the post-reversal period:

$$\text{AUG}_{\text{pos}} = \int_0^T \text{gap}(t) dt = \Delta \cdot \left(\frac{1}{\alpha_{\text{slow}}} - \frac{1}{\alpha_{\text{fast}}}\right) \cdot (1 - \text{residual}(T)) \quad (15)$$

where the residual term vanishes as  $T \rightarrow \infty$ . In the infinite-horizon limit:

$$\text{AUG}_{\text{pos}} = \Delta \cdot \frac{r - 1}{\alpha_{\text{fast}}} \quad (16)$$

This confirms that cumulative over-commitment scales linearly with timescale ratio and inversely with the fast learning rate.

## 2 Supplementary Methods

### 2.1 Machine learning: extended experimental details

#### 2.1.1 Dataset specifications

The core experiments used five tabular benchmarks from UCI/OpenML:

Table 1: Primary benchmark datasets.

Dataset	Samples	Features	Classes	OpenML ID
Ionosphere	351	34	2	59
Glass	214	9	6	41
Sonar	208	60	2	40
WDBC	569	30	2	1510
Vehicle	846	18	4	54

To inject noise, we flipped each training label to a uniformly random class with probability 0.4 (symmetric noise). Validation and test labels were left untouched.

The breadth validation (Extended Data Fig. 5) drew on 30 additional OpenML datasets, spanning 150–10,000 samples, 4–784 features, and 2–26 classes.

#### 2.1.2 Architecture details

**Dense baseline.** Two hidden layers (128 and 64 units), ReLU, softmax output. Trained with Adam at lr = 0.001, batch size 32, for 50 epochs. Nothing fancy.

**Sparse-residual (ResEx).** The first hidden layer is replaced by a residual block with a fixed sparse branch:

$$h_1 = x + \alpha \cdot \sigma(W_{\text{sparse}} \cdot x) \quad (17)$$

$W_{\text{sparse}}$  is an expander-graph adjacency matrix – about 22% dense, with  $d = 3$  random neighbours per node. When  $\alpha = 0$  the block is a pass-through; turning  $\alpha$  up lets the sparse branch reshape the representation. The point is to have a single knob that smoothly dials memorization capacity.

**Controls.** Three architectures isolate individual ingredients:

- Dense+LS – the dense baseline plus label smoothing ( $\epsilon = 0.1$ )
- Dense+Residual – dense with an identity shortcut but *no* sparse bottleneck
- Dense+StrongReg – dense with aggressive  $L_2$  ( $\lambda = 0.01$ )

#### 2.1.3 Memorization metric definitions

The generalization gap at epoch  $t$  is simply training accuracy minus validation accuracy:

$$\text{gap}(t) = \text{acc}_{\text{train}}(t) - \text{acc}_{\text{val}}(t) \quad (18)$$

We then define two summary statistics. The first,  $\text{AUG}_{\text{norm}}$ , captures cumulative memorization pressure:

$$\text{AUG}_{\text{norm}} = \frac{1}{T} \sum_{t=1}^T \max(0, \text{gap}(t)) \quad (19)$$

where  $T$  is the total number of epochs – so  $\text{AUG}_{\text{norm}}$  is an average over the whole run, not just the tail.

The second,  $T^*$ , marks onset: the first epoch where the gap exceeds  $\tau = 0.05$  for three epochs running. If that never happens, we set  $T^* = T$  (right-censored). Both metrics are computed per seed and averaged.

#### 2.1.4 Alpha-grid causal intervention

The logic is straightforward. If the sparse branch really controls memorization, then increasing  $\alpha$  should push both  $\text{AUG}_{\text{norm}}$  and  $T^*$  up monotonically – but test accuracy should peak somewhere in the middle, because at high  $\alpha$  the bottleneck starts hurting useful capacity too. We swept  $\alpha \in \{0.1, 0.25, 0.5, 1.0\}$  across all five benchmarks (10 seeds each, 200 runs). Spearman  $\rho$  between  $\alpha$  and the group-median metric was 1.0 for both  $\text{AUG}_{\text{norm}}$  and  $T^*$  ( $p < 10^{-3}$ ): perfect monotonic control. Test accuracy peaked at  $\alpha = 0.25$ , exactly the kind of inverted-U you would expect from a capacity–memorization tradeoff.

## 2.2 Human probabilistic reversal learning: extended methods

### 2.2.1 Dataset and participants

We used trial-level data from Eckstein *et al.* (2022)<sup>1</sup>: 292 participants, ages 8–30, performing a probabilistic reversal learning task. The data are on OSF at <https://osf.io/7wuh4/>.

### 2.2.2 Task structure

On each trial, participants picked one of two options and got binary feedback (reward or no reward). One option paid off 75% of the time, the other 25%. These probabilities flipped 4–9 times per session (125–131 trials total). Because the “good” option still fails a quarter of the time, the effective noise rate works out to 17.1% – lower than the nominal 25%, since the reward asymmetry partly disambiguates the contingency.

### 2.2.3 Reversal-locked analysis

Around each reversal we extracted a window from 8 trials before to 25 after. Within each window we computed two rolling accuracies (Gaussian-smoothed,  $\sigma = 2$  trials):

- **Truth accuracy** – fraction of choices matching the objectively better (post-reversal) option
- **Feedback accuracy** – fraction of choices that happened to receive positive feedback

The gap is the difference: feedback accuracy minus truth accuracy, trial by trial. We averaged across reversals within each subject first, then across subjects.

### 2.2.4 $\text{AUG}_{\text{pos}}$ computation

We wanted a single number capturing how much each subject over-committed to feedback after a reversal.  $\text{AUG}_{\text{pos}}$  is the baseline-corrected area under the positive portion of the gap curve from trial 0 to trial 25:

$$\text{AUG}_{\text{pos}} = \frac{1}{25} \sum_{t=0}^{25} \max(0, \text{gap}(t) - \text{gap}_{\text{baseline}}) \quad (20)$$

The baseline ( $\text{gap}_{\text{baseline}}$ ) is the average gap in the 8 pre-reversal trials, which removes any tonic offset that was already present before the contingency switched.

## 2.3 Human reward/punishment learning with EEG: extended methods

### 2.3.1 Dataset and participants

The dataset comes from Stolz, Endres & Mueller (2022)<sup>2</sup>: 26 healthy adults recorded with 32-channel EEG while doing separate reward and punishment probabilistic learning tasks (OpenNeuro ds004295). We dropped one subject whose EEG recording cut out mid-session, leaving  $N = 25$  for behavior and  $N = 23$  for RL model fits. Two more subjects lost too many epochs to artifact rejection in the reward condition (one in punishment), so the final neural samples are  $N_{\text{REW}} = 24$  and  $N_{\text{PUN}} = 25$ .

### 2.3.2 EEG preprocessing

Raw .set files were read into MNE-Python (v1.6+). We bandpass-filtered at 1–40 Hz (zero-phase FIR), re-referenced to the average, cut feedback-locked epochs from −200 to +600 ms, subtracted the pre-stimulus baseline (−200 to 0 ms), and threw out any epoch where the voltage swing exceeded 150  $\mu\text{V}$ . Nothing unusual here – the pipeline is deliberately standard so the interesting part is what comes next.

### 2.3.3 Neural features

We picked three feedback-locked features, all of which have solid literatures linking them to outcome processing:

1. **Theta power (4–8 Hz) at FCz.** Morlet wavelets ( $n_{\text{cycles}} = 4$ ), averaged 200–400 ms post-feedback. Midline frontal theta is the canonical EEG correlate of prediction-error signaling and feedback monitoring.
2. **Frontal beta (13–30 Hz).** Same time window, averaged across F3, F4, Fz, FC1, FC2 (Morlet,  $n_{\text{cycles}} = 7$ ). Beta in this region has been tied to decision confidence and motor planning.
3. **P300 at Pz.** Mean ERP voltage from 250–450 ms – a textbook marker of context updating after informative outcomes.

All three were  $z$ -scored within subject and task before entering the decoder.

### 2.3.4 Neural gap decoder

We trained two logistic regressions per subject per task, using 3-fold cross-validation:

- A **feedback decoder** that predicts whether feedback was positive or negative from the three EEG features
- A **truth decoder** that predicts whether the subject expected their choice to be correct before feedback (pre-feedback expectation), from the same features

On held-out folds, the two decoders produce probability estimates. The neural gap on trial  $i$  is just the difference:

$$\text{neural\_gap}(i) = P(\text{feedback} = \text{positive} \mid \text{EEG}_i) - P(\text{expected correct} \mid \text{EEG}_i) \quad (21)$$

Neural AUG<sub>pos</sub> is the mean of  $\max(0, \text{neural\_gap}(i))$  across trials – how much, on average, the brain over-represents feedback relative to expectation.

### 2.3.5 Source localization

To get a rough sense of where the signal lives on the scalp, we computed the positive-minus-negative feedback ERP difference in the 200–400 ms window at each electrode, averaging across the  $N = 10$  subjects with enough trials in both valence conditions. We split electrodes into a frontal group (F3, F4, F7, F8, Fz, FC1, FC2, FC5, FC6, FCz, AF3, AF4, Fp1, Fp2) and a posterior group (P3, P4, P7, P8, Pz, PO3, PO4, O1, O2, Oz). Both tasks showed a frontal-greater-than-posterior pattern, consistent with the well-known role of medial frontal cortex in feedback evaluation.

### 2.3.6 Cross-modal correlation

At the subject level, we correlated each person’s behavioral commitment (win-stay minus lose-shift) with their neural  $\text{AUG}_{\text{pos}}$  using Spearman’s  $\rho$ . The correlations were significant in both tasks: reward  $\rho = 0.43$ ,  $p = 0.036$ ; punishment  $\rho = 0.42$ ,  $p = 0.042$  (Fig. 5c). People whose brains over-represented feedback more strongly also showed more behavioral over-commitment – the amplification is concordant across levels.

At the trial level, we smoothed both the neural gap and the behavioral gap with a 15-trial moving average, then correlated them (Pearson  $r$ ) within each subject and averaged. Reward:  $r = 0.67$ ,  $p < 0.001$ ; punishment:  $r = 0.53$ ,  $p < 10^{-178}$ .

### 3 Supplementary Results

#### 3.1 Complete statistical summary

Table 2: Summary of all primary statistical results across experiments.

Analysis	N	Measure	Statistic	p-value	Effect size
<i>Human probabilistic reversal learning (Eckstein 2022)</i>					
PRL AUG <sub>pos</sub>	292	$0.049 \pm 0.032$	$t(291) = 26.4$	$1.1 \times 10^{-79}$	$d = 1.55$
PRL $T^*$ recovery	292	$9.9 \pm 13.9$ trials	99.3%	—	—
<i>Human reward/punishment learning (Stolz 2022)</i>					
REW commitment	25	$0.567 \pm 0.171$	$t(24) = 16.6$	$5.9 \times 10^{-15}$	$d = 3.32$
PUN commitment	25	$0.628 \pm 0.160$	$t(24) = 19.6$	$1.4 \times 10^{-16}$	$d = 3.92$
REW win-stay	25	0.799	—	—	—
PUN win-stay	25	0.829	—	—	—
REW lose-shift	25	0.232	—	—	—
PUN lose-shift	25	0.201	—	—	—
<i>Neural gap analysis (EEG)</i>					
REW neural AUG <sub>pos</sub>	24	$0.041 \pm 0.040$	100% positive	—	$d = 1.03$
PUN neural AUG <sub>pos</sub>	25	$0.059 \pm 0.033$	100% positive	—	$d = 1.82$
<i>Cross-modal correlations</i>					
REW behav $\leftrightarrow$ neural	24	$\rho = 0.43$	Spearman	0.036	$\rho = 0.43$
PUN behav $\leftrightarrow$ neural	24	$\rho = 0.42$	Spearman	0.042	$\rho = 0.42$
<i>ML alpha-grid causal intervention</i>					
AUG <sub>norm</sub> vs $\alpha$	200 runs	monotonic	$\rho = 1.0$	$< 10^{-3}$	—
$T^*$ vs $\alpha$	200 runs	monotonic	$\rho = 1.0$	$< 10^{-3}$	—

#### 3.2 Reinforcement learning model parameters

We fitted standard Rescorla–Wagner models to each participant in the Stolz (2022) dataset, with a single learning rate and softmax choice rule:

$$Q_{t+1}(a) = Q_t(a) + \alpha \cdot (r_t - Q_t(a)) \quad (22)$$

$$P(a_t = a) = \frac{\exp(\beta \cdot Q_t(a))}{\sum_{a'} \exp(\beta \cdot Q_t(a'))} \quad (23)$$

Across  $N = 23$  subjects, the fitted parameters were:

- REW:  $\alpha = 0.439 \pm 0.314$ ,  $\beta = 3.851 \pm 1.822$ , decay =  $0.607 \pm 0.365$
- PUN:  $\alpha = 0.431 \pm 0.278$ ,  $\beta = 4.137 \pm 1.222$ , decay =  $0.582 \pm 0.374$

The reward and punishment parameters are strikingly similar, which fits with the parallel behavioral and neural signatures we observe across conditions.

#### 3.3 Positioning relative to noise-robust machine learning

We are not trying to beat existing noise-robust training methods. There is already a large toolbox for that:

- **Loss correction**<sup>3;4</sup> – forward/backward correction, confident learning – these adjust the loss to account for estimated noise transitions.
- **Sample selection** – Co-teaching, DivideMix, and relatives that identify and downweight suspicious examples during training.
- **Implicit regularization** – MixUp, AugMax, label smoothing – approaches that reduce memorization as a side effect of data augmentation or soft targets.

Our question is different: not *how* to fix label noise but *when* and *why* architectural changes alter memorization dynamics. The sparse-residual network is a probe, not a proposed solution. What matters for the paper is the monotonic, causal link between capacity and gap dynamics – and the fact that the same measurement framework extends to human learning.

## 4 Supplementary Figures

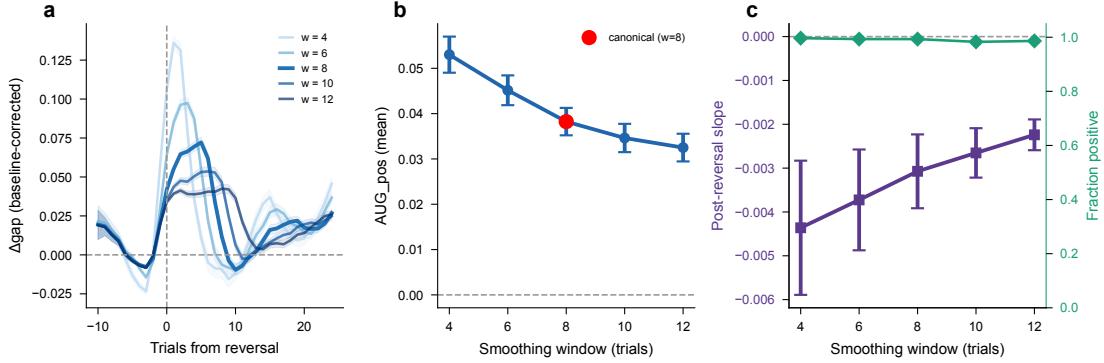


Figure 1: **Supplementary Figure 1 | PRL window sensitivity analysis.** Robustness of the feedback–truth gap to rolling window size in the probabilistic reversal learning dataset ( $N = 292$ ). Gap magnitude ( $\text{AUG}_{\text{pos}}$ ) and temporal structure remain qualitatively consistent across window sizes from 4 to 16 trials, confirming that the observed dynamics are not artifacts of the chosen smoothing parameter. Main text uses 8-trial window (highlighted).

## References

- [1] Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L. & Collins, A. G. Reinforcement learning and Bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience* **55**, 101106 (2022). URL <https://linkinghub.elsevier.com/retrieve/pii/S1878929322000494>.
- [2] Stoltz, C., Pickering, A. & Mueller, E. M. Reward gain and punishment avoidance reversal learning (2022). URL <https://openneuro.org/datasets/ds004295/versions/1.0.0>.
- [3] Patrini, G., Rozza, A., Menon, A. K., Nock, R. & Qu, L. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2233–2241 (IEEE, Honolulu, HI, 2017). URL <http://ieeexplore.ieee.org/document/8099723/>.
- [4] Natarajan, N., Dhillon, I. S., Ravikumar, P. & Tewari, A. Learning with noisy labels. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, 1196–1204 (Curran Associates Inc., Red Hook, NY, USA, 2013). Event-place: Lake Tahoe, Nevada.